Computational methods and modeling of cellular function for the optimization of protein production and the study of cellular disease states

THÈSE N^{O} 7865 (2017)

PRÉSENTÉE LE 5 OCTOBRE 2017
À LA FACULTÉ DES SCIENCES DE BASE
LABORATOIRE DE BIOTECHNOLOGIE COMPUTATIONNELLE DES SYSTÈMES
PROGRAMME DOCTORAL EN BIOTECHNOLOGIE ET GÉNIE BIOLOGIQUE

ÉCOLE POLYTECHNIQUE FÉDÉRALE DE LAUSANNE

POUR L'OBTENTION DU GRADE DE DOCTEUR ÈS SCIENCES

PAR

Joana Raquel PINTO VIEIRA

acceptée sur proposition du jury:

Prof. B. E. Ferreira De Sousa Correia, président du jury
Prof. V. Hatzimanikatis, directeur de thèse
Prof. M. Herrgard, rapporteur
Dr Y. Hashambhoy-Ramsay, rapporteuse
Prof. K. Schoonjans, rapporteuse

ÉCOLE POLYTECHNIQUE
FÉDÉRALE DE LAUSANNE
Suisse
2017

Dedication

To FFSP, (The Watchmaker)
Tic
To whom time has not been so kind to see me reach the top of this mountain;
but who grew me out of my training wheels so I could ride the path on my own.
Toc

Acknowledgements

In the words of F. Pessoa, "Success consists in being successful, not in having potential for success. Any wide piece of ground is the potential site of a palace, but there is no palace till it is built." However, no one builds palaces alone and as such, here, I would like to thank all the people who in different ways contributed to this “construction”.

I would like to start by acknowledging my thesis supervisor, Professor Vassily Hatzimanikatis, above all for his kindness and generosity. I am grateful to him for having accepted the challenge of grooming me into a researcher, for bringing me into interesting projects, and for his support and invaluable advice throughout these years.

I also thank the members of my thesis committee, Dr. Yasmin Hashambhoy-Ramsay, Professor Markus Herrgard, Professor Kristina Schoonjans and Professor Bruno Correia for their evaluation, insightful comments, and feedback on my work.

To my colleagues in LCSB, Meriç, Alex, Vikash, Tiziano, Yves, Milenko, Georgios S., Katerina, Marianne, Daniel, Arti, Jasmin, Sophia, Homa, Zhaleh, Maria, and Christine, I am thankful for their kind words, support in my work, and all the knowledge they shared with me along these years. In particular, I would like to acknowledge Anush for her fierce believe I am *Wonder Women*, Stepan for his very useful life pragmatism, Julien for all the wonderful work discussions, Noushin for her *very* useful makeup teachings but mostly for being an inspiration, Misko for being a light with his kind words and wisdom during all the (many) times I fell apart, Tuure for making my stay in Boston seem just on the other side of the street from Switzerland, Stefano for all the great laughs when we shared an office, Pierre for making me believe I can write a decent code, and Georgios F. for being the calmest person I ever met when it counts the most; but to all of them I am deeply thankful for their friendship.

My one-year internship at Merrimack Pharmaceuticals (Cambridge, MA, USA) was paramount to solidify me as a researcher. I would like to thank Vassily and Birgit for making this wonderful experience possible and all my colleagues at Merrimack, in particular, Sara G., Sara M, Omid, Florian, Tim, Jae, Andreas, Katherine, Rachel, Petra, and Daryl, for all the great moments and discussions. I am deeply grateful to Yasmin, my supervisor at Merrimack, for being a constant source of inspiration and a marvelous professional, boss, and friend.

My deepest gratitude goes to my family, in particular to my mother Fátima, my grandmother Isaura and my great-great-aunt Branca, for their love, their continuous support on my life choices, even when it meant to feel my absence, and for their unshakable belief that I have no other choice but to succeed.

Finally, I am forever indebted to Thibault, for keeping me going through the motions, the bright ups and especially the ugly downs that only sincere love can endure, and for keeping me a functional human being throughout the thesis haze.

Lausanne, August 4th, 2017

Abstract

Systems biology is a multidisciplinary field that weaves together all the basic sciences through the use of computational and bioinformatics tools, to provide a more integrative view of the complex molecular interactions taking place within and among cells. The successes in the development and improvement of techniques for high throughput –omics has rapidly increased the amount of available data. The complexity of the underlying biological system it describes requires the development of tools to properly process it and analyze it.

Computational models mathematically describe the systems interactions, allowing intrinsic properties of the data to emerge that would otherwise be overlooked. These models provide context to the data and are used to make predictions about the behavior of the system and to simulate a broader landscape of hypothesis, saving the time and cost of performing numerous experiments. However, the number of required parameters to mathematically formulate the system increases with the model complexity to integrate the available data. Thus, the development of estimation procedures and workflows to retrieve these values from literature and databases become crucial.

In this work, we developed a set of computational models, analysis tools, and pipelines to support the study of two biological systems crucial to the cell survival: metabolism and protein synthesis. Metabolism is responsible for the production of most cell biomass, including proteins. Together, these two systems balance the renewal of protein in the cell, where metabolism provides the amino acids obtained from protein breakdown to the mRNA translation machinery. Deregulation of these systems is known to cause multiple disorders, such as neurodegenerative diseases and cancer.

In the study of protein synthesis, we employ a combination of deterministic and stochastic modeling approaches to understand its intrinsic mechanistic properties and its rate-limiting steps. A better understanding of the system properties can have a

profound impact on the development of drug targets and, in particular, in the optimization of heterologous protein production. Our studies revealed that more than one factor plays a role in the speed of translation: competition for tRNA resources and the type of cognate binding interaction between tRNA and the mRNA-ribosome complex. We also derived an equation that, given the knowledge about certain intracellular parameters pertaining to the host organism of interest, can assist in the design of transcripts for optimizing heterologous protein production.

For the study of human metabolism, we established a pipeline to generate tissue-specific reduced metabolic models that can be used to study the metabolic reprogramming of different cancers and compare it with the metabolic phenotype of a healthy cell type. Despite being presented herein for human models, this pipeline is general and can be applied to the models of any organism. Starting from a human genome scale metabolic model, the pipeline improves compound annotation, identification, thermodynamics parameter retrieval, and facilitates data integration through the connection of several compound databases in a semi-automatized fashion. This work sets a standard for metabolic model assessment and curation and improves on existing tools to generate the first thermodynamically feasible reduced model of human metabolism, which is specifically tailored to the physiology and conditions under study.

Keywords: systems biology, mathematical modeling, mRNA translation, protein synthesis, stochastic simulations, ribosomes, metabolism, reduced human metabolic model, cancer metabolic reprogramming

Résumé

La biologie des systèmes est un domaine multidisciplinaire qui regroupe toutes les sciences de base grâce à l'utilisation d'outils informatiques et de bioinformatique, afin de donner une vision plus intégrée des interactions moléculaires complexes qui se déroulent à l'intérieur et entre les cellules. Les succès dans le développement et l'amélioration des techniques de haut débit des sciences “omiques” ont rapidement augmenté la quantité de données disponibles. La complexité du système biologique sous-jacent qu'il décrit nécessite le développement d'outils pour le traiter correctement et l'analyser.

Les modèles informatiques sont un outil utile qui décrivent mathématiquement les interactions des systèmes, permettant de faire émerger les propriétés intrinsèques des données qui seraient autrement ignorées. Ces modèles sont utiles pour fournir un contexte aux données, faire des prédictions sur le comportement des systèmes et simuler un paysage plus large d'hypothèses, ce qui permet d'économiser le temps et le coût d'effectuer d'innombrables expériences. Toutefois, le nombre de paramètres requis pour formuler mathématiquement le système augmente avec la complexité du modèle pour intégrer les données disponibles. Ainsi, le développement de procédures d'estimation et de flux de travail pour récupérer ces valeurs de la littérature et des bases de données devient crucial.

Dans ce travail, nous avons développé un ensemble de modèles informatiques, d'outils d'analyse et de procédures pour soutenir l'étude de deux systèmes biologiques essentiels à la survie cellulaire: le métabolisme et la synthèse des protéines. Le métabolisme est responsable de la production de la plupart de la biomasses cellulaires, y compris les protéines. Ensemble, ces deux systèmes équilibrent le renouvellement des protéines dans la cellule, où le métabolisme fournit les acides aminés obtenus à partir de la composition des protéines dans le mécanisme de traduction de l'ARNm. La dérégulation de ces systèmes est connue pour provoquer de multiples troubles, tels que les maladies neurodégénératives et le cancer.

Dans l'étude de la synthèse des protéines, nous utilisons une combinaison d'approches de modélisation déterministes et stochastiques pour comprendre ses propriétés mécaniques intrinsèques et ses étapes de limitation de débit. Une meilleure compréhension des propriétés du système peut avoir un impact profond sur le développement des cibles de médicaments et, en particulier, dans l'optimisation de la production de protéines hétérologues. Nos études ont révélé que plus d'un facteur joue un rôle dans la rapidité de la traduction: la concurrence pour les ressources de l'ARNt et le type d'interaction de liaison apparentée entre l'ARNt et le complexe ARNm-ribosome. Nous avons également obtenu une équation qui, compte tenu de la connaissance de certains paramètres intracellulaires appartenant à l'organisme hôte d'intérêt, peut aider à concevoir des transcriptions pour optimiser la production de protéines hétérologues.

Pour l'étude du métabolisme humain, nous avons établi une procedure qui génére des modèles métaboliques réduits spécifiques aux tissus qui peuvent être utilisés pour étudier la reprogrammation métabolique de différents cancers et la comparer avec le phénotype métabolique d'un type de cellules saines. En dépit d'être présenté ici pour les modèles humains, cette procedure est générale et peut être appliquée aux modèles de n'importe quel organisme. À partir d'un modèle métabolique de l'échelle du génome humain, la procédure améliore l'annotation composée, l'identification, la récupération des paramètres thermodynamiques et facilite l'intégration des données par la connexion de plusieurs bases de données composées et de manière semi-automatisée. Ce travail définit une norme pour l'évaluation et la conservation de modèles métaboliques et améliore les outils existants pour générer le premier modèle réduit de métabolisme humain thermodynamiquement réalisable, qui est spécifiquement adapté à la physiologie et aux conditions étudiées.

Mots-clés: biologie des systèmes, modélisation mathématique, traduction d'ARNm, synthèse de protéines, simulations stochastiques, ribosomes, métabolisme, modèle métabolique humain réduit, reprogrammation métabolique du cancer

Table of Contents

LIST OF TABLES ..... **XI**

LIST OF FIGURES ..... **XIII**

LIST OF ABBREVIATIONS ..... **XV**

INTRODUCTION **1**

MODELING OF COMPLEX BIOLOGICAL SYSTEMS 1

AIM & SCOPE 2

THESIS OVERVIEW 4

ARTICLES INCLUDED IN THIS THESIS 4

PART I PROTEIN SYNTHESIS OPTIMIZATION **5**

CHAPTER 1 : UNDERSTANDING THE MECHANISMS OF MRNA TRANSLATION 7

1.1 Introduction *7*

1.1.1 The role of proteins 7

1.1.2 From DNA to protein: Transcription & Translation 7

1.1.2.1 Degeneracy of genetic code 10

1.1.2.2 Synonymous codons 11

1.1.2.3 Translation and protein folding 12

1.1.2.4 Pauses and frameshifting 12

1.1.2.5 Evolutionary preserved decoding center 13

1.1.3 Deterministic modeling of mRNA translation 14

1.1.4 Biotech applications from protein synthesis modulation 16

1.2 Materials and Methods *17*

1.2.1 Modified ZH model 17

1.3 Results and Discussion *20*

1.3.1 Performance of modified ZH model 20

1.3.2 Steady state protein synthesis rate 21

1.3.3 Sensitivity analysis of modified ZH model 23

1.4 Conclusion *26*

CHAPTER 2 : ANALYSIS OF TRANSLATION ELONGATION DYNAMICS IN THE CONTEXT OF AN ESCHERICHIA COLI CELL 29

2.1 Introduction *29*

2.2 Materials and Methods *30*

2.2.1 Stochastic model of *E. coli* translation machinery 30

2.2.1.1 Translation elongation kinetics 30

2.2.1.2 Stochastic algorithm 33

2.2.1.3 Extension of stochastic framework to include tRNA abundance fluctuations 34

2.2.2 Translational resources and mRNA cell composition 35

2.2.3 Codon elongation rate 36

2.2.4 Simulation of translation in *E. coli* cell 37

2.2.4.1 Calibration of the translation system to match literature parameters 37

2.2.4.2 Heterologous expression of different Luciferase transcripts 38

2.2.5 Analysis methods for the stochastic system and parameter definitions 40

2.2.6 *In-silico* pulse-chase 42

2.2.7 Time lags of ribosome occupancy by the tRNAs 43

2.3 Results and Discussion *43*

2.3.1 General translation properties of the cell in function of growth rate 43

2.3.2 Determinants of elongation rate 45

2.3.3 Key factors on tRNA activity 50
2.3.4 Global effects of amino acid starvation and surplus in the cell 54
2.4 Conclusion *56*

PART II ESTABLISHING A PIPELINE FOR BUILDING REDUCED HUMAN METABOLIC MODELS **59**
CHAPTER 3 : PIPELINE FOR GEM PROCESSING AND INTEGRATION OF DATA AND THERMODYNAMIC PARAMETERS 61
3.1 Introduction *61*
3.1.1 Genome Scale Metabolic Networks 61
3.1.2 Constraint-based Modeling (CBM) 63
3.1.3 GEMs unification and annotation standardization 64
3.1.4 Implementation of dynamic pipeline 67
3.2 Materials and Methods *68*
3.2.1 DRAMA framework 69
3.2.2 GEM & data preprocessing 73
3.2.3 Compound thermodynamics curation 77
3.2.4 Thermodynamics-based Flux Balance Analysis (TFA) 79
3.2.5 GEMs used in this chapter 80
3.2.6 Physiological mammalian/human cell parameters 80
3.3 Results and Discussion *81*
3.3.1 Mapping compounds of GEMs 81
3.3.2 Reaction balance and assessment of pre-assigned directionalities 85
3.3.3 Thermodynamics curation gain by automatizing 88
3.4 Conclusion *89*
CHAPTER 4 : FROM HUMAN GEMS TO CONSISTENTLY DERIVED REDUCED HUMAN METABOLIC NETWORKS 93
4.1 Introduction *93*
4.1.1 Cancer studies with genome scale models 93
4.1.2 Consistent decrease of GEM complexity tailored to system under study 97
4.2 Materials and Methods *98*
4.2.1 Recon 2 preprocessing 98
4.2.2 Data & parameter integration into Recon 2 98
4.2.3 Reduction of GEMs to data-driven networks 102
4.3 Results *104*
4.3.1 Generating data-driven reduced models 104
4.3.1.1 Assessment of dataset variability 104
4.3.1.2 Minimal media and assessment of extracellular reactions (uptakes/secretions) 105
4.3.1.3 GEM assessment with data for specific physiology 106
4.3.1.4 Selection of core subsystems for reduction 107
4.3.1.5 Reaction thermodynamics coverage per subsystem in the GEM 110
4.3.2 *RedHuman* assessment & validation 110
4.4 Conclusion and path forward *118*

CONCLUSION **121**

APPENDIX A SUPPLEMENTARY TEXTS AND METHODS **123**
A.1 SUPPLEMENTARY TEXTS AND METHODS FOR CHAPTER 1 123
A.1.1 Modified ZH model equations *123*
A.1.2 Reproduction of experiments with modified ZH model *128*
A.1.3 Sobol's Method for GSA *128*
A.2 SUPPLEMENTARY TEXTS AND METHODS FOR CHAPTER 2 129
A.2.1 Cell composition in ribosome and tRNA molecules *129*
A.2.2 mRNA sequences present in the simulated cell *130*
A.2.3 Derivation of full deterministic equation for computation of codon elongation rate *132*
A.3 SUPPLEMENTARY TEXTS AND METHODS FOR CHAPTER 3 138
A.3.1 Database web services protocols & system requirements *138*
A.3.2 Databases in local server *138*
A.4 SUPPLEMENTARY TEXTS AND METHODS FOR CHAPTER 4 139

A.4.1 Studies with GEMs *139*

APPENDIX B SUPPLEMENTARY FIGURES **143**

B.1 SUPPLEMENTARY FIGURES FOR CHAPTER 1 143

B.2 SUPPLEMENTARY FIGURES FOR CHAPTER 2 144

B.3 SUPPLEMENTARY FIGURES FOR CHAPTER 3 161

B.4 SUPPLEMENTARY FIGURES FOR CHAPTER 4 162

APPENDIX C SUPPLEMENTARY TABLES **169**

C.1 SUPPLEMENTARY TABLES FOR CHAPTER 1 169

C.2 SUPPLEMENTARY TABLES FOR CHAPTER 2 172

C.3 SUPPLEMENTARY TABLES FOR CHAPTER 3 182

C.4 SUPPLEMENTARY TABLES FOR CHAPTER 4 192

BIBLIOGRAPHY **215**

CURRICULUM VITAE **239**

List of Tables

Table 2.1 Statistics on mean ribosome occupancy time lags and total number of events per decoding 52
Table 3.1 List of databases accessed within GEMap. 74
Table 3.2 Mammalian cell physiological parameters. The pH values were obtained from (168). The 81
Table 3.3 Statistics on the mapping of Recon 2 v4 compounds using GEMap at mode 1 by searching..... 84
Table 3.4 Statistics on reaction directionality at different stages of flux constraining for Recon 2 v4 88
*Table 4.1 Summary of human normal (NH) datasets types collected per tissue. The * indicates data 100*
*Table 4.2 Summary of human cancer (CH) datasets types collected per tissue. The * indicates datase.....101*
Table 4.3 List of aggregated datasets to build different metabolic tissue phenotypes for model reduc.....102

Table C.1.1 Rate constants for the ribosomal kinetic pathway during translation elongation obtain 169
Table C.1.2 aa-tRNA concentrations per binding interaction for the growth rate 0.4h^{-1}. Individual.....170
Table C.2.1 Rate constants for the ribosomal kinetic pathway during translation elongation obtain 172
Table C.2.2 List of wobble pairing nucleotides for each tRNA species derived from the list of codons 173
Table C.2.3 List of tRNA species that are cognate WC, cognate WB and near-cognate to each codon. 174
Table C.2.4 aa-tRNA concentrations per binding interaction estimated for the growth rate 1.07h^{-1} 177
Table C.2.6 Comparison between mRNA sequencing data from E. coli at low and high growth rates 178
Table C.2.7 Detailed mRNA Luciferase transcripts used for recording ribosome occupancy time lag 179
Table C.3.1 Summary of Web services provided per database and possible query results of interest. 182
Table C.3.2 Reactions that present $\Delta rG'°$ sign switched due to a change in ionic strength value fro.....182
Table C.3.3 Listing of compounds that did not have common external database identifiers between 183
Table C.3.4 Statistics on reaction directionality at different stages of flux constraining for Recon 2 190
Table C.3.5 Statistics on thermodynamics curation per metabolite and reaction for two GEM netw.....191
Table C.4.1 Detailed listing of dataset sources for normal and cancer phenotypes. In the info field, ...192
Table C.4.2 Information on growth rate measurements and data extraction methodology (data po 198
Table C.4.3 Detailed listing of information on conversion factors and assumptions made to convert.....203
Table C.4.4 List of uptakes blocked in Recon 2 v4 based on the presence of Coenzyme A (CoA), acyl c 208
Table C.4.5 List of essential amino acids and other metabolites found in mammalian cell minimal m.....209
Table C.4.6 List of subsystems from Recon 2 v4 that are not explicitly kept in the RedHuman core ne.....211
Table C.4.7 List of metabolites that have lost extracellular reactions in RedHuman (green), that hav 212
Table C.4.8 The 15 lumped reactions added to the core network of RedHuman to generate a networ.....214

List of Figures

Figure 1.1 Main steps in protein synthesis. The process depicted here is for a Eukaryote organism *8*
Figure 1.2 The three main steps of the translation process: initiation, elongation and termination. *9*
*Figure 1.3 a) Escherichia coli transfer tRNA carrying the amino acid Valine (*tRNAUACVal*). Deg* *11*
Figure 1.4 Schematic representation of our modified ZH model, which includes initial selection a *18*
Figure 1.5 The blue and red curves represent the cumulative amount of cognate and near cognat *20*
Figure 1.6 Comparison of dipeptide formation for cognate (blue) and near cognate (red) with an *21*
Figure 1.7 Optimal protein synthesis rate at different stages of ribosomal density for four E. coli. *22*
Figure 1.8 a) Main effect and b) total effect indices computed using Sobol's method for the evalu *24*
*Figure 1.9 Main (*Si*) and total (*STi*) effects of modified ZH model computed with respect to the p* *24*
Figure 1.10 a) Main effect and b) total effect indices computed with Sobol's procedure with resp *25*
Figure 1.11 Correlation between the aa-tRNA concentrations of cognate, near-cognate and non- *26*
Figure 2.1 Schematic representation of the ribosome kinetics of the translation elongation cycle *32*
*Figure 2.2 (a) Distribution of protein synthesis rate (*Vp*) for the different growth rates. Red bar* *46*
Figure 2.3 (a) Comparison between simulated (sim) and experimental time-evolution curves of *47*
*Figure 2.4 (a) Main and total effects (*Ski, STki*) on the value of* ν*r due to a change in ribosome* *50*
Figure 2.5 Number of tRNA molecules of species i *active in translation (*tRNAia*) in function of* *53*
Figure 2.6 Each marker represents the relative change of mean elongation rate of the cell's tran *55*
Figure 3.1 Constraint based modeling (CBM) formulation. A metabolic network is converted into *62*
Figure 3.2 Schema of DRAMA main working pipelines for GEM annotation, data integration and *69*
Figure 3.3 Visualization of format and content of DRAMAcore repository, here presented as a MA *71*
Figure 3.4 Summarized procedure for data integration into the model. Datasets are selected thr *72*
Figure 3.5 GEMap pipeline is shown here with the search in mode 1 where the purpose is to mat *75*
Figure 3.6 GEMap pipeline is shown here with the search in mode 2 oriented to KEGG compound *76*
Figure 3.7 Schema of DRAMA pipeline integrating GEMap results (or any other available comp *78*
Figure 3.8 a) Comparison of metabolite mapping coverage in Recon 2 v4 between GEMap using *85*
Figure 3.9 Statistics on number of reactions in Recon 2 v4 that were balanced, that required pr *85*
Figure 3.10 Number of reactions in Recon 2 v4 whose original flux constraint bounds were modi. *87*
Figure 3.11 Assessment of network thermodynamics constraints coverge for Recon 2 v4 and iM *89*
Figure 4.1 Complete pipeline for processing and reducing a human GEM for the study of cancer *96*
Figure 4.2 Overview of DRAMA full pipeline applied to Recon 2 v4 GEM. The GEM is mapped with *100*
Figure 4.3 Summary of redGEM workflow for the systematic, semi-automatized reduction of GE *103*
Figure 4.4 Illustration of pipeline for generation of reduced models with redGEM. Different redu *104*
Figure 4.5 Fraction of metabolic reactions per subsystem (excluding transport reactions) in Reco *111*
Figure 4.6 Comparison of metabolic network size between Recon 2 v4 and RedHuman. *112*
Figure 4.7 Quantification of reactions in the RedHuman network with respect to the original Rec *114*
Figure 4.8 Maximum growth yield for each (blue bars) RedHuman with AllNormal and AllCancer *117*

Figure B.1.1 Range of total codon occupancies computed with modified HZ model for E. coli gene ya *143*
*Figure B.2.1 The total number of ribosomes per cell (*RT*) (squares) and the total number of tRNA m* *144*
Figure B.2.2 The rate of mRNA synthesis per cell (circles) was obtained from (58) for the growth rat *144*
Figure B.2.3 Visual comparison between the E. coli codon usage frequency (CU) and the mRNA codo *145*
Figure B.2.4 (a) Time evolution of free ribosome amount. At simulation time 0s the number of free *146*
Figure B.2.5 Time evolution of the number of free tRNA molecules of each species for an E. coli cell *147*
Figure B.2.6 (a) Mean distance between ribosomes in function of growth rate. Results from comput *148*
Figure B.2.7 In-silico pulse chase experiment curves obtained by averaging the time-evolution of th *149*
Figure B.2.8 Translation time profiles for each of the seven Luciferase transcripts studied. Black ver *150*
Figure B.2.9 (a) Fold change in protein translation from all transcripts relative to WT. (b) Elongati *151*

*Figure B.2.10 Elongation rate (*υr*) of an mRNA species in function of the change of each of the ribos*.....*152*
Figure B.2.11 Sensitivity of ribosome kinetic parameters (a) measured at in vitro conditions at 37ºC *153*
Figure B.2.12 Mean ribosome occupancy time lags Δtbidsj *per decoding stage* ds *= (A-site OFF, A-sie* *154*
Figure B.2.13 Total number of events Nbidsj *per decoding stage* ds *= (A-site OFF, A-site PROOF, P-sit*.....*155*
Figure B.2.14 Interaction-based mRNA codon usage frequency (IBmCU) displayed for each tRNA spe.....*156*
Figure B.2.16 Relative deviation of average elongation rate from all mRNA species in the cell at 1.0h.....*157*
Figure B.2.17 Relative deviation of average codon elongation rate (computed with keff*) from all c*.....*158*
Figure B.2.18 For each tRNA species in the cell, the keff *values of all its cognate codons (WC and/or*.....*159*
Figure B.2.19 mRNA codon elongation rate (mCU) presented per codon species and tRNA isoacceptor.....*160*
Figure B.3.1 Distribution of the deviation scores computed as ΔrG'°I = 0.15-ΔrG'°I = 0.25ΔrG'°I.....*161*
Figure B.4.1. Distribution of intracellular metabolite levels measured for the 60 cancer cell lines in N.....*162*
Figure B.4.2 Distribution of intracellular metabolite levels measured for the 2 prostate cancer cell lin *163*
Figure B.4.3 Distribution of intracellular metabolite levels measured for the 6 leukemia cancer cell li.....*164*
Figure B.4.4 Comparison metabolomics and fluxomics datasets merged to produced normal vs cancer *165*
Figure B.4.5 Number of metabolic reactions per subsystem (excluding transport reactions) in Recon 2 .....*166*
Figure B.4.6 Survey of first human reduction attempt departing from Recon 2 v4 and using fluxomics.....*167*
Figure B.4.7 a) Combination of intracellular concentration values for argininosuccinate, fumarate a *168*
Figure B.4.8 The reduction procedure in redGEM and lumpGEM (205, 206) can be adapted be perfor.....*168*

List of Abbreviations

DNA	Deoxyribonucleic acid
RNA	Ribonucleic acid
mRNA	Messenger RNA
tRNA	Transfer RNA
Met-tRNA$_i^{Met}$	Methionyl-initiator transfer RNA
E. coli	*Escherichia coli*
UTR	Untranslated region
ZH model	Reference to Zouridis and Hatzimanikatis model {Zouridis:2008kh}
Aa-tRNA:EF-Tu:GTP	Ternary complex (denoted as aa-tRNA throughout the text for simplicity)
IC	Initiation complex
EF-Tu	Prokaryote translation elongation factor Tu
EF-G	Prokaryote translation elongation factor G
GSA	Global Sensitivity Analysis
DRAMA	Data Repository for Automatic Metabolomics Administration
GEMap	Genome scale metabolic model compound mapping
GEM	GEnome-scale metabolic model
KEGG	Kyoto Encyclopedia of Genes and Genomes
ChEBI	Chemical Entities of Biological Interest
HMDB	The Human Metabolome Database
HMR	Human Metabolic Atlas
EHMN	Edinburgh Human Metabolic Network
GPR	Gene-Protein-Reaction relationship
CBM	Constraint-Based Modeling
COBRA	COnstraint-Based Reconstruction Analysis toolbox
FBA	Flux Balance Analysis
TFA	Thermodynamics-based Flux Balance Analysis
FVA	Flux Variability Analysis
TFVA	Thermodynamics-based Flux Variability Analysis
GCM	Group Contribution Method
LP	Linear Programming
MILP	Mixed Integer Linear Programming
SBML	Systems Biology Markup Language
SMILES	Simplified molecular-input line-entry system
InChI	IUPAC International Chemical Identifier

Introduction

Modeling of complex biological systems

Systems Biology is an interdisciplinary field of study that attempts to model complex biological systems by applying physics and chemistry concepts. Since these systems are a representation of many complex biological interactions taking place in a cell or tissue, it is common to take bioinformatics approaches to simulate and analyze them. The aim of such complex system-wide representations is to understand the functionality and behavior of the biological systems under study, thus promoting the discovery of targets for drug development and molecular biomarkers for the presence of disease states.

The innovation and improvement in high-throughput -omics techniques in the recent past has increased the size, scope, and complexity of the data that needs to be analyzed. The development of computational modeling concepts, frameworks, and standard-operation procedures, as well as bioinformatics tools, are becoming increasingly crucial for the processing of large amounts of data and its integration into even more complex computational models. These models allow the analysis of the physiological system under a multiplicity of conditions, the test of hypothesis, as well as the ability to make reliable predictions. Good computational tools can assist the design of experiments and, more importantly, provide an efficient screening of different hypothetical scenarios, by rapidly evaluating and testing different conditions in *in-silico* experiments, saving the time and cost of many laborious experimental procedures.

Computational models can be used to explore biological systems at different scales, spanning from local chemical interactions inside and across organelles of the most basic living block, the cell, to interactions between cells of the same or even

different tissues. Independent of the scale, it is the level of detail one wishes to examine these systems, as well as the assumptions one is willing to make, that will dictate their complexity and number of intervening biological compounds.

Aim & Scope

This work focuses on the modeling of protein synthesis and cellular metabolism, both key systems for cell survival and maintenance.

Protein synthesis is the cellular process responsible for the production of all proteins and, subsequently, many enzymes necessary for the regulation of crucial cellular functions. Perturbations to the delicate balance of this machinery are the basis of many diseases, such as neurodegenerative diseases and cancer, among others (1). Problems can occur at different stages of protein synthesis (2-4). Mutations of the mRNA strands can lead to differences in the final sequence of the protein, which can affect both co- and post-translation folding processes, impairing it from achieving its final functional structure. Transcription and translation factors assist in different steps of the protein synthesis process. Mutations in any of the genes encoding these molecules will interfere with the mechanism and alter patterns of mRNA expression and protein levels. A deeper knowledge of the mechanisms of protein synthesis will contribute to better understand and target the system modifications leading to disease.

The understanding of the system mechanism allows for the use of its properties to fine tune protein synthesis for industrial applications. For instance, optimization of the production of recombinant protein therapeutics in cultures of mammalian cells is a topic of continuous research (5). Identification of the rate limiting steps of translation in the context of the protein synthesis machinery and resources in the host organism can guide the design of more efficient expression vectors. Other ways to improve protein titer levels have focused on optimization of the culture media of the host cells and/or genetically modifying the host cells to promote growth. We focus on the translation elongation mechanism of protein synthesis and, through the development and extension

of deterministic and stochastic algorithms, we aim to explore the overall system dynamics and its rate limiting steps, i.e., the kinetic rates that exert the most control over the rate of protein production.

Metabolism in a cell comprises a set of biochemical reactions operated by metabolic enzymes that are regulated according to the cell requirements and function. Catabolic pathways are a set of metabolic reactions responsible for the breakdown of molecules into their constitutive building blocks units, which are then reassembled through anabolic pathways into new molecules (proteins, nucleotides, lipids, ...) that constitute the cell biomass. Production of these molecules is regulated and depends on cellular function and growth stage.

Cellular metabolism can be mathematically represented with Genome Scale Models (GEMs) that contain all the metabolic network reactions known for an organism based on a listing of annotated genes that encode for metabolic enzymes. GEMs have become increasingly popular in biotech industry and systems medicine with applications in host cell engineering for optimized protein production and in the study of many metabolic diseases.

The application of GEMs in systems medicine has become quite successful, in particular for cancer studies (6-9), after the appearance of the first human genome scale metabolic reconstructions (6, 10-12) and assisted by the devolvement of algorithms to integrate high throughput -omics data, such as transcriptomics, proteomics, fluxomics and metabolomics, (13-18).

In this work, we establish a pipeline to assist in the annotation of GEMs for reaction thermodynamics curation and data integration and improve on existing tools and workflows to generate reduced data-driven phenotypic models that can be used in the study of cancer metabolic reprogramming. Throughout this work, we aim to provide a workflow description that is fully reproducible and can be used by non-experts in the field.

Thesis overview

This thesis is divided in two main parts:

- Part I focuses on the modeling of protein synthesis, in particular, the process of translation elongation. In these studies, we perform a global analysis of translation in the context of an *Escherichia coli* cell and determine its rate-limiting steps with the purpose of assisting on optimization of transcript design.

- Part II focuses on the description of an implemented pipeline for GEM processing and [re]curation in terms of metabolite annotation and compound structure, with the purpose of facilitating data and thermodynamics parameter integration. This pipeline establishes the complete procedure from GEM to the first thermodynamically consistent derived reduced human metabolic network.

Articles included in this thesis

The following list of articles and their publication status is included in this thesis:

J. Vieira, J. Racle, and V. Hatzimanikatis (2016) *Analysis of Translation Elongation Dynamics in the Context of an Escherichia coli Cell*. Biophysical Journal

Status: *Published*

See Part I, Chapter 2: Analysis of translation elongation dynamics in the context of an *Escherichia coli* cell

J. Vieira, M. Masid, A. Chiappino-Pepe, V. Pandey, M. Ataman, and V. Hatzimanikatis (2017) *RedHuman*: a reduced human metabolic network for thermodynamics-based flux balance analysis of cancer physiology

Status: *In preparation (provisory title)*

See Part II for pipeline description.

Part I Protein Synthesis Optimization

Chapter 1: Understanding the mechanisms of mRNA translation

1.1 Introduction

1.1.1 The role of proteins

Protein synthesis plays an important role in biological systems since its products constitute most of the molecular machinery required for cell regulation, growth, and functionality. Proteins are sequences built by combining amino acids in a specific order from a selection of a total of 20 amino acids species. These amino acid species are conserved across organisms and are said to be essential if the organism cannot produce them, having thus to include it in its diet. The electrochemical interactions between the amino acids forming the protein sequence determine its final 3D structure, which has to acquire a precise shape to fulfill its biological function in the cell. A slight deviation from its correct structure may impair the protein in its interaction with cell receptors for signaling purposes, in its role as a catalyzer of a chemical transformation, or even in forming a complex with other proteins.

The process of protein synthesis is very complex and encompasses many stages with many intervening enzymes and molecules. When not functioning properly it can lead to cell death or disease. Protein folding errors are commonly associated with neurodegenerative diseases, such as Alzheimer's, and cancer. Besides errors in sequence, deregulation at different stages of the protein synthesis machinery can lead to overproduction of mutated proteins that promote cancer progression and survival (see (1) for a review).

1.1.2 From DNA to protein: Transcription & Translation

The process of protein synthesis starts with the *Transcription* step where a gene sequence in a DNA strand is copied into a template called a messenger RNA (mRNA), as depicted in Figure 1.1. The mRNA sequence contains the same information of the DNA

coding region but it is stored in a base-pair complementarity to the DNA strand. The RNA polymerase is the enzyme that catalyzes the transcription process with the assistance of transcription factors that facilitate its binding to, progress along and release from the DNA strand yielding a complete mRNA template. The mRNA template is exported from the cell nucleus, in the case of a Eukaryote organism, and may undergo an intermediate step of splicing where the introns (non-protein coding regions) are removed from the sequence and only the exons (protein coding regions) remain.

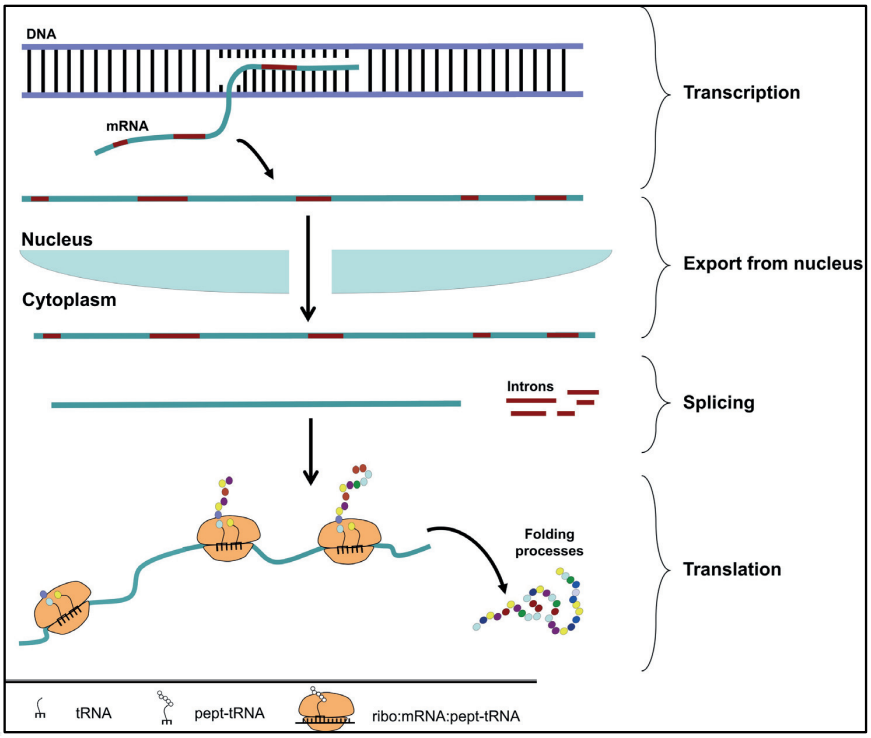


Figure 1.1 Main steps in protein synthesis. The process depicted here is for a Eukaryote organism with transcription occurring in the nucleus and post-transcript splicing.

The second main step in the process of protein synthesis, which is the main focus of the work that will follow in Chapters 1 and 2, is *Translation*. During translation, the mRNA strand is decoded and its corresponding polypeptide chain is synthesized by an enzyme called the *ribosome*. Many ribosomes can simultaneously translate an mRNA strand forming a *polyribosome*. The amino acids are transported to the ribosome by another type of RNA molecule, the transfer RNA (tRNA). When tRNAs are charged with

an amino acid they are called amino-acyl tRNAs (aa-tRNAs). In the last stage of protein synthesis, other processes and enzymes are activated to ensure the correct folding of the protein to acquire its functional 3D structure.

There are three main steps in translation (Figure 1.2). In the *initiation step*, it is formed a pre-initiation complex consisting of the methionyl-initiator tRNA (Met-$tRNA_i^{Met}$) bound to a specific location on the small ribosome subunit. This pre-initiation complex then binds to the 5' end of the mRNA strand and proceeds to the scanning of the 5' untranslated region (UTR) until it recognizes the start codon (AUG), which signals the beginning of the open reading frame coding the protein. At this stage, the large ribosomal subunit assemblies with the preinitiation complex bound to the mRNA and forms the complete initiation complex. Several initiation factors, which are heavily regulated in eukaryotes, are required for the formation of the initiation complex (IC).

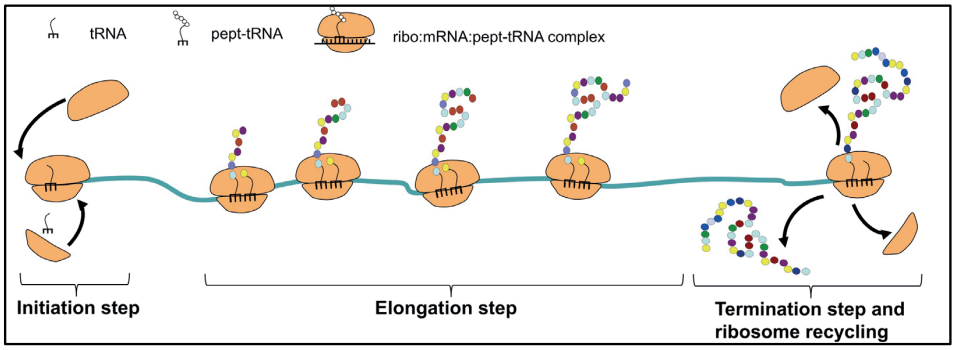


Figure 1.2 The three main steps of the translation process: initiation, elongation and termination.

During the *elongation step*, the ribosome decodes the mRNA open reading frame and builds the polypeptide chain by catalyzing the reactions involved in the selection and recognition of aa-tRNAs transporting the correct amino acid. Several ribosomes can be simultaneously active translating one mRNA chain (polyribosome or polysome). The length of these polyribosomes contributes to the protein synthesis rate in an organism and are dictated by the translation initiation rate of the mRNA, the existence of space allowing for a new pre-initiation complex to bind, and the speed of elongation that moves forward the ribosome *traffic jam*. The synthesis of the polypeptide chain is complete at the *termination step* when the ribosome reaches one of the possible stop

codons (UAG, UAA, or UGA). At this stage, the polypeptide chain is hydrolyzed from the ribosome-decoding center and the ribosome enters the stage of *ribosome recycling* where the large and small ribosome subunits dissociate and unbind from the mRNA, thus becoming available for another translation round.

1.1.2.1 Degeneracy of genetic code

An amino acid is coded by every three nucleotides (codon) in the mRNA sequence. Each aa-tRNA has an anticodon region in its sequence that recognizes the codon in the mRNA strand (Figure 1.3a). However, there is a degeneracy that allows for different codons with mismatches on the third nucleotide position (the wobble position) with respect to the tRNA anticodon to be recognized by the same aa-tRNA (Figure 1.3b). Although the codons are recognized during translation of the mRNA, these Wobble binding interactions between the codon in the mRNA and the anticodon of the aa-tRNA are weaker than the ones resulting from a perfect match. The ribosome catalyzes these reactions at a slower rate, influencing the speed of translation elongation. Evidence for these differences in elongation speed can be found in the literature. For instance, Sorensen and Pedersen (19) found that the two codons GAA and GAG that are decoded by the same aa-tRNA present a 3.4-fold difference in elongation rate.

Based on the mismatches between the codon and aa-tRNA anticodon four levels of interactions can be defined:

- Watson-Crick cognate: all three nucleotides match between codon and anticodon;
- Wobble cognate: there is a mismatch of wobble type in the third position between codon and anticodon;
- Near-cognate: there is one mismatch between codon and anticodon that does not involve the wobble position;
- Non-cognate: there is more than one mismatch between codon and anticodon that do not involve the wobble position.

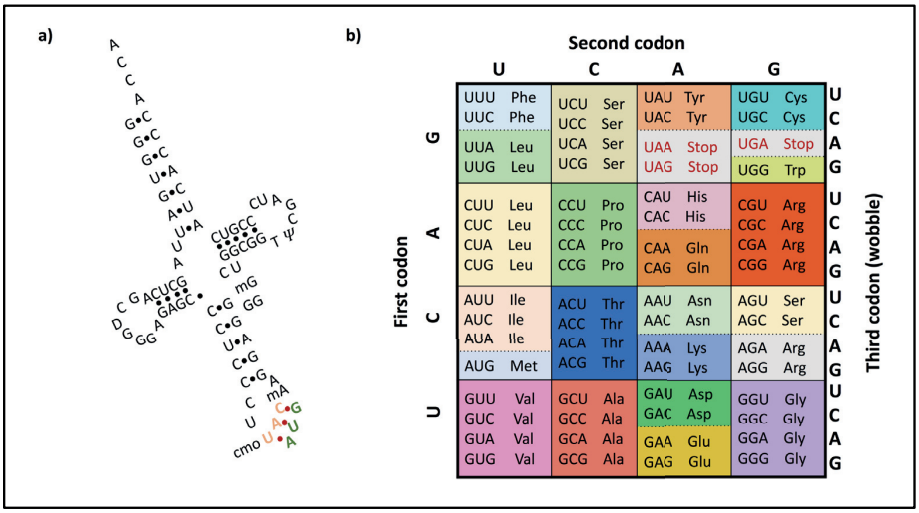


Figure 1.3 a) *Escherichia coli* transfer tRNA carrying the amino acid Valine ($tRNA^{Val}_{UAC}$). Degeneracy allows this tRNA to also decode the codons GUA, GUG and GUU. b) Table representing the degeneracy of the genetic code: the amino acids and the respective degenerated codons that encode them.

1.1.2.2 Synonymous codons

The codons that are recognized by the same aa-tRNA (in the wobble position) are known as synonymous codons because, if replaced in the mRNA sequence, the polypeptide sequence resulting from its translation would be the same. Despite these wobble interactions being weaker and affecting the speed of elongation of these synonymous codons, there are studies that indicate the existence of other levels of modulation in elongation speed.

On one hand, Varenne and colleagues (20) and Curran and Yarus (21) found that the speed of translation elongation of different codons depends on their correspondent aa-tRNA availabilities. The premise behind this is that the slower translating codons, also known as rare codons, stall the ribosome while they wait for the low abundant cognate aa-tRNAs to bind. Individual aa-tRNA abundances are different among organisms and, within the same organism, they depend on physiological conditions and on tissue functionality. Studies involving synonymous codon substitution of rare codons have shown that these codons are associated with ribosome pausing and their replacement by more frequent ones was observed to decrease protein specific activity (22, 23), which can be associated with protein misfolding, as different structural domains require different speeds to be formed (24).

In the other hand, Bonekamp and colleagues (25) measured the elongation rates of 12 codons in *Escherichia coli (E. coli*) and showed that they do not always correlate with their cognate aa-tRNA abundances or codon usage. These differences were postulated to arise either from a slower recognition of the codon-anticodon formation due to differences in base pairing, such as the wobble position, or as an effect of ternary complex competition.

1.1.2.3 Translation and protein folding

The protein is required to be folded in a certain configuration (native state), which is absolutely necessary in order for it to be functional. It is currently accepted that the protein starts to fold already during translation ((22) and references cited therein). Studies on *S. cerevisiae* and *E. coli* genes have shown that replacing rare codons by frequent ones reduced pauses in translation and increased the rate of protein synthesis, but the resulting protein presented lower specific activity that could be associated with misfolding (26, 27).

Thus, it seems that the different codons along the mRNA strand encode more than the simple amino acid sequence of the protein. The different elongation rates resulting from ribosome pausing and tRNA abundance allow the necessary time in order for certain structural properties of the protein start emerging, assisting it towards its final conformation. Furthermore, it appears that different elongation speeds in rare codons encode for different structural domains. For instance, faster-forming subdomains would require less time to fold than bigger domains (28).

1.1.2.4 Pauses and frameshifting

Regulatory mechanisms can be in the origin of pauses in translation that slow down protein synthesis. One such example, that involves the formation of mRNA secondary structure, is the case of translation of *E. coli* trp operon. This mRNA has secondary structures that are formed in different sequence locations that are regulated by the levels of tryptophan present in the cell. These secondary structures affect the movement of the ribosome and result in the attenuation of translation elongation (29).

Ribosome pauses during translation can lead to frameshifting events where the ribosome adjusts its position in the mRNA strand leading to an alternative reading of the coded sequence and hence to a different protein sequence. Different mechanisms can originate frameshifting events. For instance, a -1 frameshift event results from a ribosome slippage over one nucleotide in the 5' direction of the mRNA sequence. Jacks and colleagues (30) have observed in the replication of the Rous sarcoma virus that the formation of mRNA secondary structure in front of the ribosome drives ribosomal translocation from one codon to the next, but in the opposite direction.

A +1 frameshift event has been observed as a result of a pause in translation elongation during the decoding of a rare codon with low available cognate aa-tRNA abundance. This is likely to occur if the aa-tRNA bound to the previous codon has also an affinity to the codon formed by shifting one nucleotide in the 3' direction of the mRNA (31). Sundararajan and colleagues (32) also showed that a +1 frameshift event can be induced by a different mechanism involving the binding of a near-cognate aa-tRNA in the absence of cognate aa-tRNA that stimulates the binding of the next aa-tRNA with an alternate frame through interactions with its wobble position.

1.1.2.5 Evolutionary preserved decoding center

The protein synthesis machinery differs between prokaryote and eukaryote organisms. One such difference is the coupling of transcription and translation in prokaryotic cells, i.e., the mRNAs start to be translated before their transcription is completed, whereas the mRNAs in eukaryotic cells are transcribed in the nucleus from where they are exported into the cytosol for the translation machinery to take over (Figure 1.1). The existence of this coupling in bacteria allows for another level of regulation where the rate of translation elongation controls the elongation rate of transcription through a cooperative mechanism between ribosome and RNA polymerases, as observed in (33).

However, since the work presented in this thesis focus specifically on the study of translation elongation, we are more interested in the differences between organisms at that level, which will influence the assumptions we make for generalizations. Translation is a complex process that involves a chain of reactions in all its three main steps (initiation, elongation and termination/ribosome recycling) and an informative

comparison of the different translation steps between prokaryotes and eukaryotes is presented in (34). These reactions depend on several translation factors (proteins) that differ in their coding sequence and in their function across organisms (prokaryotes, eukaryotes, and archaea). The advances in the understanding of translation from the use of tools such as cryo-electron microscopy and X-ray crystal structures (35, 36) led to the identification of many such translation factors and helped to elucidate their function.

Interestingly, it has been observed that the elongation step across organisms is the only translation step for which homologous factors have been identified across different organisms (37). In fact, several studies in *E. coli* and *S. Cerevisiae* indicate that the ribosome decoding center, which is responsible for the accuracy of translation, share identical ribosomal domains in both organisms (38). Indeed, the homologous elongation factors play an important role in the recognition of codon and aa-tRNA anticodon complementarity, in the peptide bond formation and in the translocation of the ribosome from one codon of the mRNA to the next, which are reactions that occur at the conserved ribosome-decoding center.

These findings constitute a solid basis for the assumption that the mechanisms of translation elongation are the same, or very similar, in prokaryotes, eukaryotes, and archaea. However, the same assumption does not hold for the initiation, termination and ribosome recycling steps.

1.1.3 Deterministic modeling of mRNA translation

There have been some efforts in the past to investigate the mechanism of translation with deterministic models that take into account the kinetics of the ribosome and its movement along an mRNA, transitioning from one codon to the next. These models are described by a system of ordinary differential equations and they compute the average response from a population of cells, which indeed reflects the majority of the experiments available, where data is collected from a population of cells thus ignoring the stochasticity present at the individual level. However, there has been recently an emergence of single-cell experiments which will further assist in the development of stochastic models.

Gibbs and colleagues (39, 40) presented a mathematical model of protein translation where the mRNA molecule is represented by a 1-D lattice, with each lattice site corresponding to one codon (3 nucleotide residues). The ribosome moves one codon at a time, while occupying several codons at the same time. Later, Heinrich and Rapoport (41) refined this model and they performed a computational study from which they concluded that the initiation *and* elongation phases of translation determine the rate of protein synthesis under normal cell conditions. They also observed in this study that when the ribosomes are distributed uniformly along the mRNA the termination rate is fastest under physiological conditions. Having Heinrich's and Rapoport's model as a base, Mehra and Hatzimanikatis (42) performed a genome-wide study of translation networks where they investigated the influence of adding extra mRNA transcripts on the synthesis rate of individual mRNAs.

The underlying assumption of these models is that all the different codons have the same rate of elongation. However, different studies (mentioned before) found the rate of translation elongation of individual codons to be dependent on factors such as the abundance of aa-tRNAs and their anticodon binding affinity to the codons in the mRNA. Gilchrist and colleagues (43) proposed a model for the translation of one mRNA strand that accounted for differences in the elongation rate of individual codons assuming that these were proportional to the corresponding aa-tRNA abundance in the cell. Furthermore, the model also included steps for ribosome recycling and for the occurrence of nonsense errors due to frameshifting, false termination or premature ribosome release. Despite this improvement, this model did not take into account the polysome formation and their effect on elongation rate. Later models (44-46) account for both the polysome formation and the differences in the elongation rate of individual codons by either assuming proportionality with the cognate aa-tRNA abundance (44, 46) or by estimating elongation speed based on the proportionality between cognate and competitor near-cognate tRNAs (45).

Taking advantage of the recent understanding of ribosome kinetics (35-37) and the bulk rapid-mixing kinetic experiments for the *E. coli* translation system *in vitro* (47-49), Zouridis and Hatzimanikatis (50, 51) formulated a codon-specific mathematical model for translation elongation (ZH model) that accounts for i) polysome formation, ii) cognate tRNA abundance and non-cognate competition, and iii) the known intermediate

kinetic steps of the ribosome for peptide bond formation reported in (49) combined with the kinetic mechanism of ribosome translocation, i.e., ribosome movement from one codon to the next, based on the work of Savelsbergh and colleagues (52).

1.1.4 Biotech applications from protein synthesis modulation

All efforts that contribute to building upon the knowledge of the mechanisms of translation, bring us closer to formulate strategies for the modulation of those processes in the cell. This is relevant for many Biotech companies, and in particular pharmaceutical companies, that have an interest in maximizing the yield of recombinant proteins and/or designing drugs to target translation deficiencies.

In recombinant protein production, host cell systems are engineered to efficiently express a gene of interest such that the respective protein is secreted from the system in high amounts for potential use in clinics, research, or other applications. Briefly, the system is designed so that the expression of the gene of interest is dependent on the expression of a gene essential for the host cell organism. The host cells are cultured with a drug that blocks the expression of the essential gene inducing an amplification on its expression to balance the effect of the blocking drug, and, as a consequence, amplifying the expression of the gene of interest. A more complete understanding of the determinants of translation elongation can assist in the design of expression vectors for the protein of interest such that its translation occurs more rapidly. This is indeed the focus of the work presented in Chapter 2.

1.2 Materials and Methods

1.2.1 Modified ZH model

Several studies (53, 54) report the existence of an induced fit mechanism where the ribosome assesses the binding affinity between the aa-tRNA and the mRNA codon at ribosome A site through a series of substrate-induced conformational changes. Selection and recognition of the cognate aa-tRNA is processed in two consecutive steps: initial selection and proofreading. These two selection steps are necessary for efficient aa-tRNA discrimination and they contribute individually with similar efficiencies for the overall selectivity (49).

We started from the ZH model (50) and further extended it to include the two aforementioned selection steps (Figure 1.4). For binding with the ribosome, the aa-tRNA binds to EF-Tu:GTP in a reaction catalyzed by the elongation factor EF-Ts forming the ternary complex (TC). For simplicity, throughout the text, TC and aa-tRNA are used interchangeably, unless otherwise stated. The ZH model accounted for TC competition solely by assuming that all near- and non-cognate aa-tRNA species bind to the ribosome at the codon-independent binding site (k_1, k_{-1}) without further progression. However, the error frequency measured *in vivo* for *E. coli*, i.e., a measure of the incorporation of wrong aa-tRNAs, of 6×10^{-4} (55) and 5×10^{-3} on internal mRNA codons (56), indicates that the near-cognate aa-tRNAs can proceed to, and, in rare occasions, beyond the proofreading step.

In our modified ZH model (Figure 1.4), we discriminate the binding of the near- and non-cognate TCs to the ribosomal A site: the affinity of the near-cognate aa-tRNA is assessed in both steps of selection, whereas the non-cognate aa-tRNAs are assumed not to pass beyond the initial selection step since codon recognition that triggers GTP hydrolysis does not occur in the relevant physiological time (47).

After the codon-independent binding to the ribosome (k_1) it follows the recognition step (k_2) where, if the codon is recognized, GTPase activation of EF-Tu elongation factor is triggered (k_3), otherwise it can be rejected through the reverse reactions of initial selection. Conformational rearrangements of EF-Tu:GTP to EF-Tu:GDP are accompanied by the release of inorganic phosphate (k_4). The elongation factor EF-Tu

loses its affinity for the aa-tRNA and dissociates from the ribosome with rate constant ($k_{release}$).

From this stage of the kinetic pathway, our model allows for proofreading step. If the aa-tRNA is recognized as the correct anticodon it accommodates to the 50S ribosomal A site (k_5) triggering peptide bond formation, whereas an incorrect aa-tRNA will be rejected from the ribosome (k_{rej}). Peptide bond formation is followed by the ribosome translocation, i.e., an EF-G-dependent displacement of the tRNA-mRNA complex that was previously on the A site onto the P site with respect to the ribosome, leaving the A site available for aa-tRNA binding and decoding of the next mRNA codon. Since the translocation intermediate steps seem to be codon independent we assume that the kinetic pathway is the same for both cognate and near-cognate aa-tRNA binding. The ribosome translocation takes place at state 9, where a new codon is positioned in the A-site.

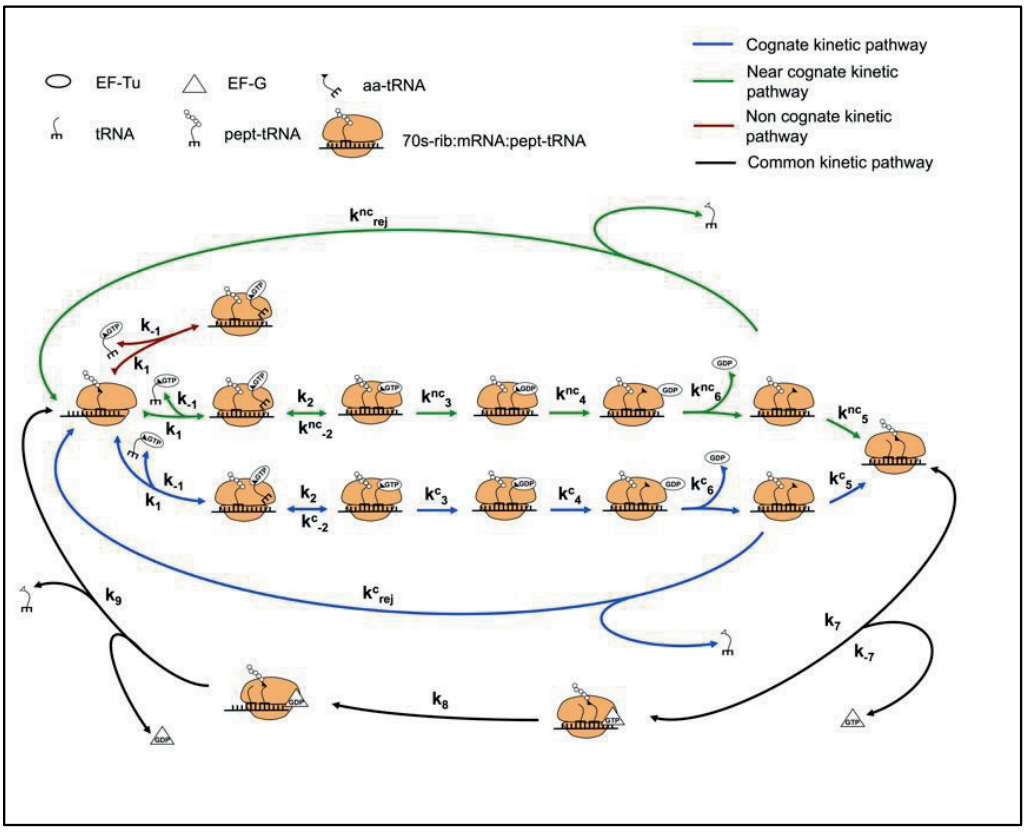


Figure 1.4 Schematic representation of our modified ZH model, which includes initial selection and proofreading kinetics for near-cognate TCs.

The model equations representing the ribosome kinetics in Figure 1.4 are detailed in section A.1.1. Table C.1.1 contains the kinetic rate values and their definitions for the *E. coli* ribosome. These were estimated from experiments where bulk rapid-mixing techniques were used at conditions that reproduce the overall aa-tRNA selectivity observed *in vivo* for *E. coli*. We note that, since less intermediate steps were used in the original ZH model, we introduced a discontinuity in the index of the ribosome states in our modified ZH model to establish a connection with subsequent model extensions of the ribosome kinetics in Chapter 2.

As described in (50), the model can be formulated with a system of equations

$$\frac{d\boldsymbol{x}}{dt} = \boldsymbol{N} \cdot \boldsymbol{V}(\boldsymbol{x}, \boldsymbol{p}) \tag{1.1}$$

where $\boldsymbol{N}$ is the stoichiometric matrix, $\boldsymbol{x}$ is a vector containing the fractional occupancies of each codon j representing the probability of having a codon occupied by the A site of a ribosome, $\boldsymbol{p}$ refers to the parameters of the model (translation kinetic rates and cellular concentrations of ribosomes, aa-tRNAs and translation factors), and $\boldsymbol{V}$ is the vector of the reaction fluxes representing translation initiation, elongation and termination and given by

$$\begin{cases} V_I = k_I \cdot R^f \cdot W_I \\ V_j = k_{eff}^j \cdot x_j, \quad j = [1, \dots, n-1], \\ V_T = k_T \cdot x_n \end{cases} \tag{1.2}$$

The term W_I represents the probability that another ribosome can bind to the first codon on an mRNA strand given that a bound ribosome occupies L_R codons

$$W_I = 1 - \sum_{j=1}^{L_R} x_j. \tag{1.3}$$

This model determines the protein synthesis rate in function of an effective elongation rate (k_{eff}^j) for each codon j along the mRNA transcript (see Eq. A.1.4 and section A.1.1 for steady state derivation), which is a function of the ribosome kinetic rate constants for elongation and the concentrations of elongation factors, cognate and competitor aa-tRNAs.

1.3 Results and Discussion

1.3.1 Performance of modified ZH model

We compared the simulation results from the modified ZH model (Figure 1.4) to the experimentally measured levels of GTP hydrolysis in (49). To reproduce the experimental conditions, we simulate the translation of a mRNA segment of the form auguuuuuu(...)uaa (26 UUU codons in between the start and stop codons) with 0.2 μM of either cognate or near cognate TC so that no competition takes place, and 2.8 μM of initiation complexes at 20°C in high fidelity conditions. For the near cognate simulation, the codon UUU is replaced by CUC. The dynamic simulations (see section A.1.2) were performed using the ribosome kinetic rates in Table C.1.1. The simulated curves fit well with the experimental data points (Figure 1.5) indicating that the model formulation up to GTP hydrolysis (k_3) is representative of the system behavior.

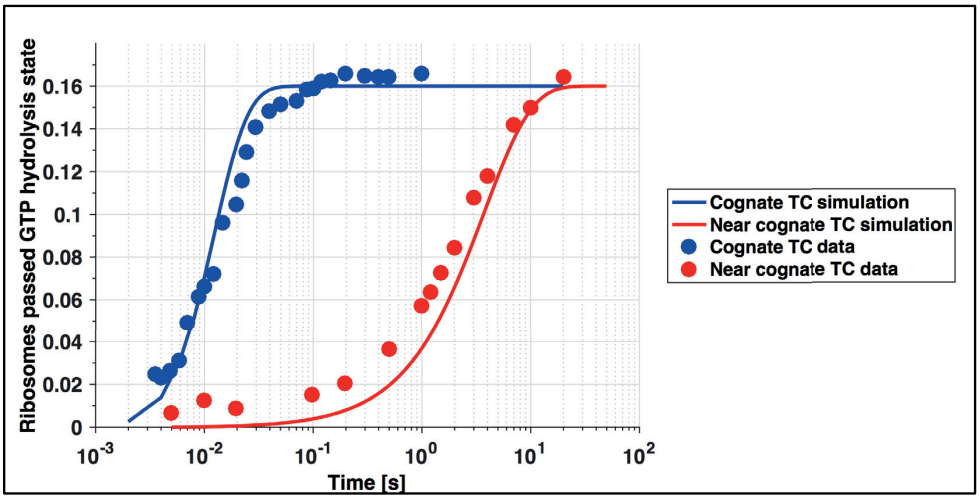


Figure 1.5 The blue and red curves represent the cumulative amount of cognate and near cognate TCs hydrolyzed over time, respectively. Experimental points were extracted from (49) using a Matlab script.

We further investigated how the experimentally measured levels of dipeptide bond formation compare to our modified ZH model (Figure 1.6). Experiment and simulation were performed as described above for GTP hydrolysis case (see section

A.1.2), except for the modified amount of IC used (1 µM). The simulated amount of dipeptide bond formation in time for the recognition of cognate TC is delayed with respect to the data. Modifying the model to exclude the release of EF-Tu:GDP from the ribosome ($k_{release}$) before accommodation / peptide bond formation (k_5) produces a better fit. This extra reaction step, which is present in the original ZH model, introduces a delay in the dipeptide formation. However, its removal is in agreement with the literature (53), which represents this step as a parallel event to the accommodation. We note here that the GTP hydrolysis rate, which is comparable to the accommodation rate, limits the near cognate dipeptide reaction and hence, the removal of EF-Tu:GDP release step does not influence the fit for the near cognate case.

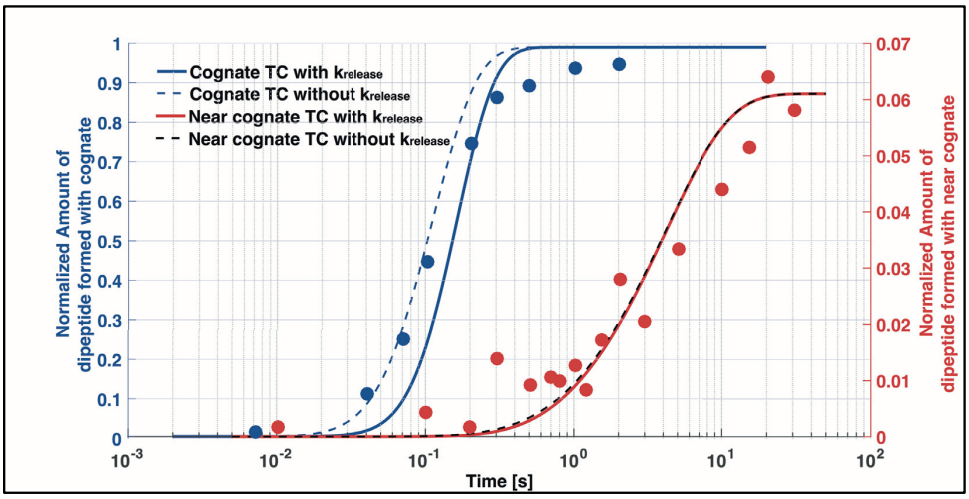


Figure 1.6 Comparison of dipeptide formation for cognate (blue) and near cognate (red) with and without the GDP release sep ($k_{release}$) in the model reaction chain. Experimental points were taken from (49) using a Matlab script.

1.3.2 Steady state protein synthesis rate

The *protein synthesis rate per mRNA* (also referred to as *elongation rate*) was computed by applying an optimization framework developed by Racle and colleagues (57) to the modified ZH model. This optimization allows for the determination of the optimal protein synthesis rate and polysome configuration given a fixed concentration of initiation, elongation, and termination factors, and a fixed pool of ribosomes and aa-tRNAs. As an example, Figure 1.7 displays the elongation rate determined by the

optimization framework for four different *E. coli* mRNA transcripts with different lengths in function of the ribosomal density (ρ). The ribosomal density is a measure of the ribosome crowding along the mRNA transcript and is defined as

$$\rho = \frac{\sum_{j=1}^{n} x_j}{n} \cdot P = \frac{P \cdot L_R}{n}, \tag{1.4}$$

where P is the polysome size, i.e., the number of ribosomes bound to mRNA transcript, n is the number of codons in the mRNA transcript and L_R is the length of the ribosome in terms of number of codons it occupies on the mRNA.

The simulations were performed using ribosome and aa-tRNA concentrations determined for *E. coli* growing at 0.4h^{-1}. The free ribosome concentration (R^f) was computed by assuming that 80% of the total ribosome concentration is active in translation (58). The total ribosome concentration at 0.4h^{-1} was obtained from (58) and the individual aa-tRNA concentrations at 0.4h^{-1} used for the calculation of cognate, near- and non-cognate aa-tRNA concentrations for each codon (Table C.1.2) were obtained from (59).

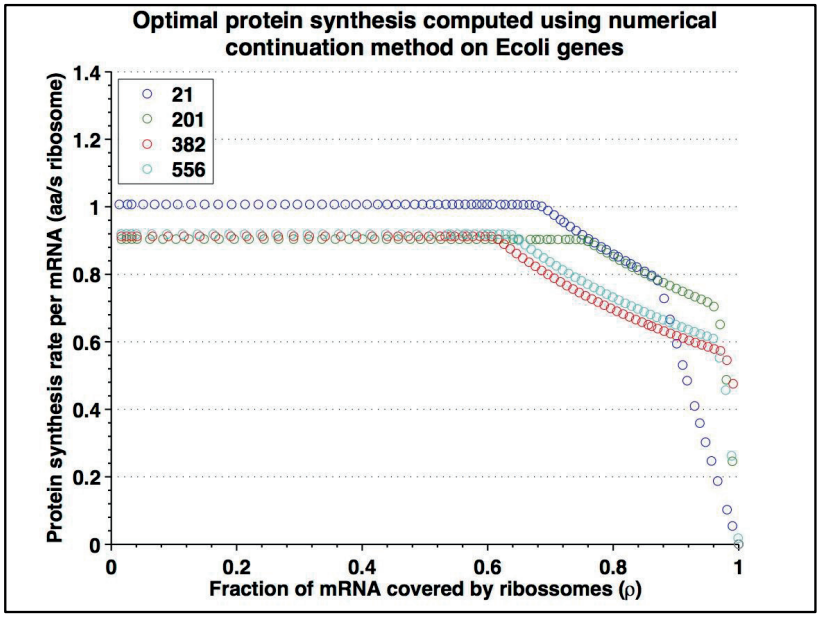


Figure 1.7 Optimal protein synthesis rate at different stages of ribosomal density for four *E. coli* mRNA transcripts with increasing length.

1.3.3 Sensitivity analysis of modified ZH model

We use the Sobol's method (see section A.1.3), a global sensitivity analysis (GSA) procedure, to evaluate the influence of the ribosome kinetic rates in the response of the modified ZH model. We selected the *E. coli* gene yahD (mRNA transcript with 302 codons length) and computed main effects (Eq. A.1.12) and total effect (Eq. A.1.13) of each kinetic rate constants on different model responses.

To assess the parameter sensitivity during one codon elongation, we simulated the time course of peptide bond formation for the binding of a cognate aa-tRNA to the codon following the start codon in this transcript. The individual kinetic rates contribute differently to the time course of peptide bond formation, as shown by the sensitivities computed at each time point (Figure 1.8). As expected, the initial phase is mostly controlled by both the initial selection step, where the aa-tRNA binds to the ribosome, and the codon:anticodon recognition steps. However, the rates of phosphate release (k_4) and accommodation (k_5) present an overall high contribution that becomes more important in the vicinity of the dipeptide bond formation stage. The understanding of the influence of different kinetic rates at different time points in a chain of reactions can be useful for assigning weights in parameter estimation procedures. In this particular case, this analysis informed us about the relevance of k_4 in the kinetics of dipeptide bind formation, which, interestingly, is a parameter with high associated uncertainty due to its difficult determination from experiments.

To assess the parameter sensitivity on the translation of a complete mRNA sequence, we simulated the translation of the *E. coli* gene yahD for a sample of elongation kinetic rates (Figure B1.1) and for a translation initiation rate fixed to a low value (to avoid ribosome crowding interferences). The main and total effects for the modified ZH model with respect to the protein synthesis rate (Figure 1.9) show (in agreement with the results from Zouridis and Hatzimanikatis (51)) that aa-tRNA competition, determined by the binding and codon recognition rates (k_{-1}, k_2, and k_{-2}^{nc}), does play a major role in the modulation of the speed of translation elongation, followed by a minor influence of the rate of cognate peptide band formation. The rate constants related to the translocation and the rejection by proofreading play a minor role in the total codon occupancies. However, we note that these results were generated for a

translation state limited by initiation, which means that ribosome crowding is not expected to play a role.

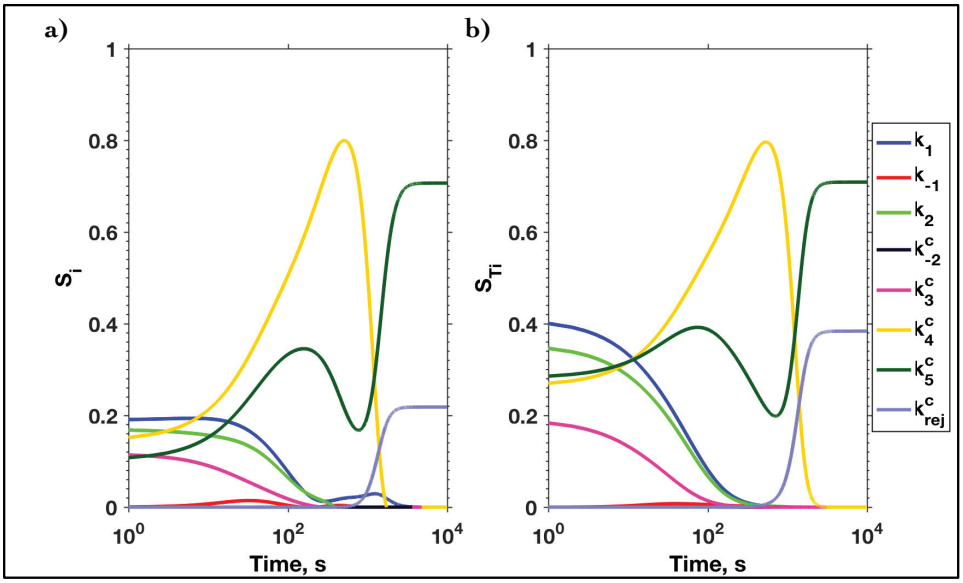


Figure 1.8 a) Main effect and b) total effect indices computed using Sobol's method for the evaluation of the critical kinetic rates during the time course of peptide bond formation for a cognate aa-tRNA.

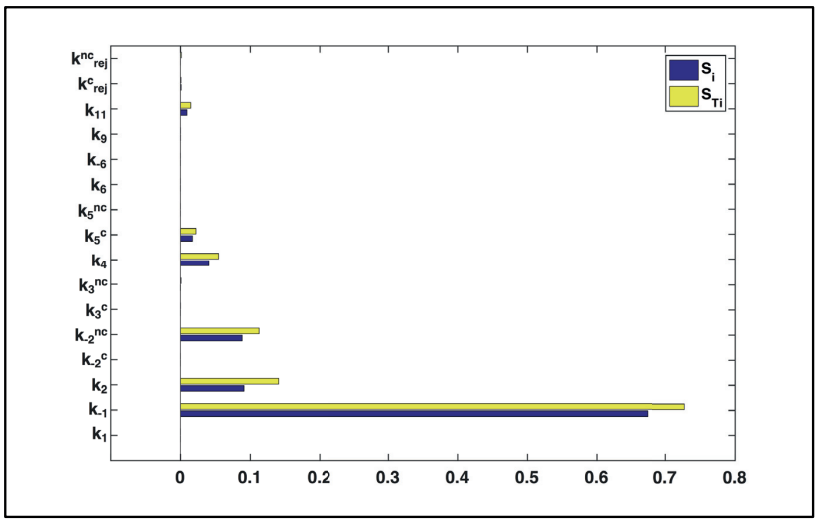


Figure 1.9 Main (S_i) and total (S_{T_i}) effects of modified ZH model computed with respect to the protein synthesis rate.

To further study the codon-dependent sensitivity along the mRNA sequence we looked into the changes in the total codon occupancies along the mRNA transcript. The total codon occupancies (x) are the defined in section 1.2.1 (Eq. 1.1 and 1.2) and represent the probability of having a codon occupied by the A site of a ribosome, serving as a measure of the polysome size. In agreement with the previous results, the main and total effects with respect to the changes in the total codon occupancies (Figure 1.10) show (k_{-1}) as a dominant influential parameter. Interestingly, k_{-2}^{nc} and k_2 present "bursts" of opposing sensitivity: when one increases the other decreases (see selected zoom in Figure 1.11). These results are consistent with a model that is heavily influenced by competition between TCs binding to a ribosome, which is highlighted by the dominance of the reverse reaction in codon-independent binding step (k_{-1}). In particular, the end region of the "bursts" of sensitivity for parameter k_{-2}^{nc} correlate with codons that present a higher concentration of near-cognate aa-tRNA, decreasing the speed of elongation in the upstream codons.

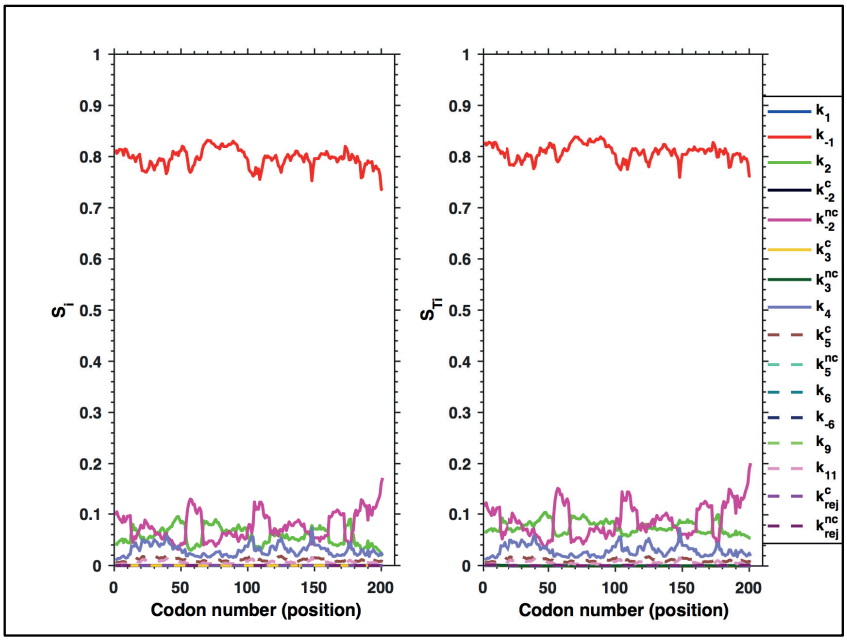


Figure 1.10 a) Main effect and b) total effect indices computed with Sobol's procedure with respect to the total codon occupancies.

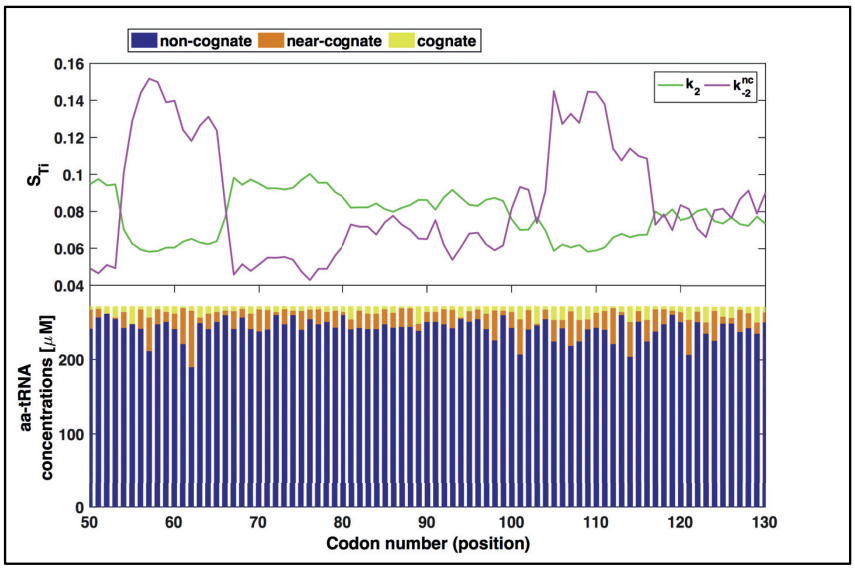


Figure 1.11 Correlation between the aa-tRNA concentrations of cognate, near-cognate and non-cognate species of a codon and the codon sensitivity indices with respect to the total codon occupancy.

1.4 Conclusion

Understanding the mechanisms of translation and its rate limiting steps is crucial for both the development of drug targets and improvement of heterologous protein production with many biotechnological applications, such as in pharmaceutical and biofuel industries.

Despite many advances in the knowledge of the ribosome structure and function, there is still much discussion around the determinants of translation elongation with experiments and computational studies disagreeing, in particular, between the influence of the cognate aa-tRNA abundance and aa-tRNA competition. However, Spencer and colleagues (60) have recently observed that the discrimination between cognate aa-tRNAs with Watson-Crick (*WC*) and wobble (*WB*) binding interactions plays a major role in the modulation of codon the elongation rate.

We have started from the ZH model and extended it to include the initial selection and proofreading stages for the near-cognate TC binding. Our studies have shown that the modified ZH model is able to reproduce the experimental measurements

of the levels of dipeptide formation and corrected for a ribosome kinetic step that was overlooked in the original ZH model ($k_{release}$). Also, in agreement with the analysis from the ZH model, our sensitivity analysis results indicate that aa-tRNA competition is the major factor influencing elongation rate.

The developed models do not have a detailed kinetic description that discriminates between cognate *WC* or *WB* types such that the influence of the interplay between the binding of those aa-tRNAs on the modulation of the elongation rate can be studied. Furthermore, if aa-tRNA levels (cognate or competitor) do play a role in translation, it is more relevant to perform studies where the actual amount of free aa-tRNA is taken into account for the binding with the ribosome and not its total amount in the cell. This amount of aa-tRNA available for binding is dependent on the number of active ribosomes and mRNA sequences being translated.

Our studies of translation continue in Chapter 2, where we have extended a stochastic model of translation elongation based on previous work. This model is able to dynamically track of the levels of free ribosomes and aa-tRNAs, the position of bound ribosomes in the mRNA sequences, and the aa-tRNA species bound to the different ribosomal sites (A, P and E). The translation elongation kinetics of this model was also extended to discriminate between cognate *WC* and *WB* binding interactions.

Chapter 2: Analysis of translation elongation dynamics in the context of an Escherichia coli cell

2.1 Introduction

Despite the advances in the knowledge of ribosome kinetics, such as the unveiling of the structure and function of several ribosomal domains with cryo-electron microscopy and X-ray crystallography (35, 36), or with the development of bulk rapid-mixing kinetics and single-molecule experiments for the study of ribosome reaction kinetics and the dynamics of translation events (61-63), the dynamics of the translation process and its rate-limiting steps are still not completely understood. The availability of cognate tRNA for a codon is generally accepted as the determinant of translation elongation rate. However, studies of computational or experimental character that attempt to identify the rate-limiting steps of translation have not been able to provide a consensus on this matter. A computational study of translation using a mechanistic model has found that the competition between cognate tRNAs and nonspecific binding tRNAs (near-cognate (*nc*) and non-cognate (*non*) tRNAs) is the rate-limiting step in translation (51). Another computational study has identified specifically the competition between cognate and near-cognate tRNAs as the determinant in translation rates (64). More recently, in a computational model that does not take competition into account, the concentration of ternary complex aa-tRNA:GTP:EF-Tu was found to limit elongation rate (65). In two recent experimental studies involving synonymous codon replacements, the key factor in translation elongation rate was attributed to the tRNA availability in somewhat different ways. Spencer and colleagues (60) showed that the determinant of codon translation modulation is the availability of cognate tRNA with Watson-Crick vs. wobble interactions, whereas Rosenblum and colleagues (66) suggested that cognate tRNA abundance is the key factor.

Recent stochastic models (67, 68) enable the study of the translation dynamics for an organism's representative set of mRNA sequences allowing for the study of more complex dynamics, such as ribosome crowding effects and the dynamics of tRNA availability in a whole-cell context, which is more difficult to address with deterministic

models. These stochastic models take into account the fluctuations on the availability of tRNA and ribosomal resources, however, despite their complexity, they still do not provide a complete kinetics for competition. Using a stochastic framework, we simulate the translation process based on the available ribosome kinetics as determined in (53, 69-71), which describes fully the tRNA competition and differentiates between a cognate Watson-Crick (*WC*) and a cognate wobble (*WB*) tRNA binding interaction. We simulate the simultaneous translation of a representative pool of *Escherichia coli* mRNA sequences under a range of different growth rates for which the number of ribosomes and the concentrations of each tRNA species are known. We show that two distinct mechanisms modulate the speed at which each codon is translated: (i) the amount of competitor tRNA and (ii) the type of cognate binding interaction (*WC* vs. *WB*), which combined optimize elongation rate of a heterologous transcript added to the cell. Formulating the translation process deterministically by extending on previous work (51), we derive an equation that estimates the codon elongation rates based on the amount of free competitor and cognate (*WC*, *WB*) tRNAs. We compare the predictions of this equation with the ones from our stochastic model, and we show its potential to assist on the design of optimized heterologous transcripts by synonymous codon substitution.

2.2 Materials and Methods

2.2.1 Stochastic model of *E. coli* translation machinery

2.2.1.1 Translation elongation kinetics

The ribosome kinetics for translation elongation cycle of each codon of an mRNA sequence was obtained from *in vitro* experiments detailed in Table C.2.1 and is represented schematically in Figure 2.1. The four different kinetic pathways (cognate Watson-Crick (*WC*), cognate wobble (*WB*), near-cognate (*nc*), and non-cognate (*non*)) represent the different types of tRNA binding to the mRNA-ribosome complex.

The binding interaction is defined as cognate Watson-Crick (*WC*) if the three nucleotides of the codon match the three nucleotides of the tRNA anticodon, whereas it is defined as cognate wobble (*WB*) when the first two nucleotides of the codon match the last two nucleotides of the tRNA anticodon and the third nucleotide of the codon forms a wobble base pair with the first nucleotide of the tRNA anticodon. A near-cognate binding interaction is defined by a nucleotide mismatch in one of the three base pairs, whereas a non-cognate binding interaction is defined by either a nucleotide mismatch in two of the three base pairs or in all three base pairs. Note that in our definition of near- and non-cognate binding interactions we take into account the possible wobble interactions, i.e., if a wobble interaction is present on the third base pair, the binding is defined as near-cognate (*nc*) when there is a mismatch on the first or second base pair, and it is non-cognate (*non*) when there are mismatches of both the first and second base pairs. This definition is in agreement with the work of Rodnina and colleagues (53) from where we assembled the kinetic rate constants for the ribosome elongation steps.

The ribosome kinetics for *WC*, *nc* and *non* interactions are obtained from (53) at 20°C. The ribosome kinetics for *WB* was obtained from (70) at 20°C, where cells not expressing tRNA-Ala2 had its matching codon GCC decoded by the isoacceptor tRNA-Ala1B via a wobble binding interaction. Based on these biochemical assays the translation elongation cycle is divided in two stages where the codon-anticodon match is evaluated resulting in the possibility of rejecting the tRNA: the initial selection stage (states 1-3) and the proofreading stage (states 5-1). After the tRNA accommodation and peptide bond formation from states 5 to 6, the ribosome kinetics for the mRNA-tRNA translocation between states 6 and 11 is a combination of rate limiting steps from (71) at 25°C and the remaining steps are from (69) at 37°C (fast rate constants that do not limit the system). During the step at which the translated codon and tRNA are shifted to the P-site (state 9 to 10), the next codon of the mRNA sequence is placed at the A-site for decoding and at the same time the deacylated-tRNA that was previously at the P-site is translocated to the E-site.

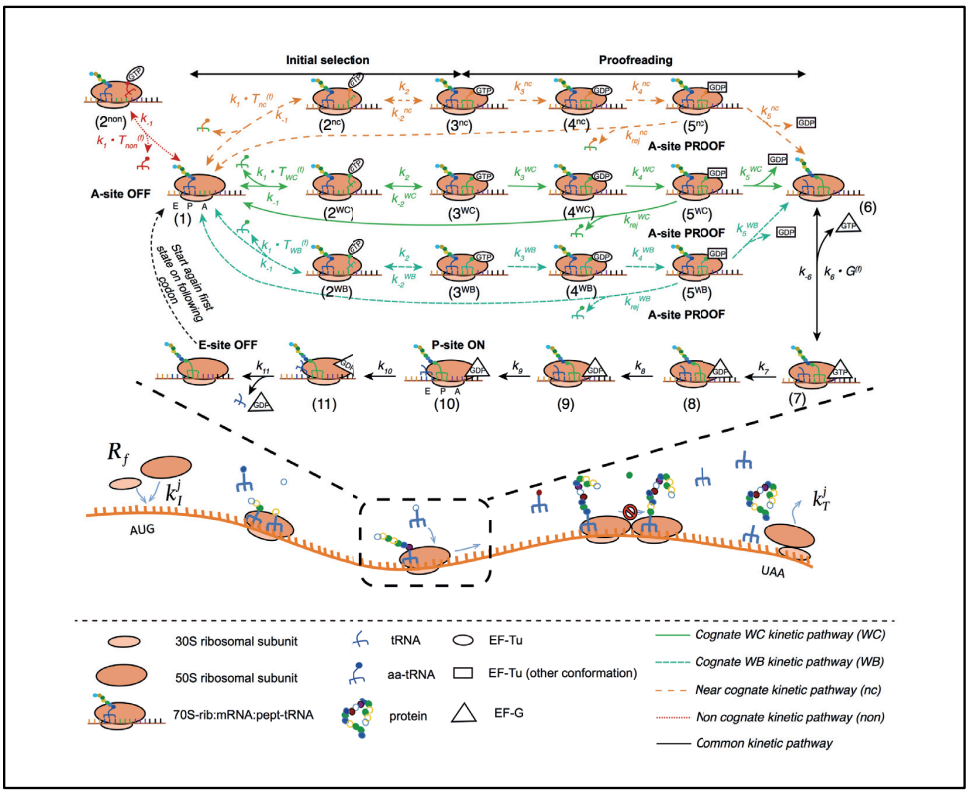


Figure 2.1 Schematic representation of the ribosome kinetics of the translation elongation cycle during which a polypeptide is synthesized following the decoding of its corresponding mRNA sequence. The four pathways represent the different types of codon-anticodon interaction (*WC*, *WB*, *nc*, *non*). After the tRNA accommodation and peptide bond formation from state 5 to state 6 the subsequent kinetic pathway is assumed to be common for the different types of binding interactions, as the kinetic steps no longer depend on the codon-anticodon recognition. $T_{WC}^{(f)}$, $T_{WB}^{(f)}$, $T_{nc}^{(f)}$ and $T_{non}^{(f)}$ are the concentration of free cognate *WC*, cognate *WB*, near-cognate and non-cognate tRNAs, respectively, for the codon being translated. A-site OFF, A-site PROOF and E-site OFF correspond to the positions where tRNA is released from the mRNA-ribosome complex and P-site ON corresponds to the position where ribosome translocation to the next codon occurs and hence the tRNA in the A-site is placed in the P-site and the one in P-site is placed in the E-site.

In our model, we assume that: (i) the ribosome kinetics for *WB* binding interaction is the same independently of the type of wobble mismatch and codon involved; and (ii) the translocation kinetics is common to the *WC*, *WB* and *nc* binding interactions, as the kinetic steps no longer depend on the codon-anticodon interaction.

Furthermore, the kinetics of the tRNA charging with an amino acid and binding with EF-Tu, mediated by EF-Ts, is not taken into account in the model. We instead assume that the tRNA is instantaneously recharged after leaving the ribosome E-site and

that the finalized ternary complex aa-tRNA:GTP:EF-Tu is readily available for binding with the ribosome, and hence not limiting translation. This assumption is consistent with the observation that 90% of EF-Tu is estimated to be present in the form of the ternary complex (72). We also note that at steady state the rates of uncharged tRNA degradation (>120 min in (73)) and aa-tRNA degradation (100-1000h in (74)) occur in a much longer time scale than the time scale for the cycle of elongation until the tRNA is released from the E-site (see Table 2.1 in section 2.3.3). This leads to the rate of tRNA charging being equal to the rate of tRNA release from the E-site and to the rate of aa-tRNA binding to the A-site. For simplicity, tRNA throughout the text denotes the finalized ternary complex aa-tRNA:GTP:EF-Tu ready to bind to the mRNA-ribosome complex.

2.2.1.2 Stochastic algorithm

Stochastic models can follow the evolution of large complex systems, tracking each of its components and providing a distribution of solutions for the model. Recent single-cell experiments indicate there is a high heterogeneous cellular behavior, which justify the use of stochastic modeling approaches.

We simulated the dynamics of translation with an exact continuous time stochastic algorithm (75) based on previous work (76). Briefly:

- The simulations of the system describing the *E. coli* cell translation machinery are initialized by setting all the ribosomes to a free state and unbound from the mRNA sequences;
- At each iteration, the next elongation reaction step is randomly selected and the time step for that reaction to occur is evaluated;
- The identity of the ribosome and the mRNA undergoing the reaction is randomly selected from a subset that is ready to take that reaction step;
- The propensities of the reactions, i.e. the probability of their occurrence as the next step, is updated based on the reaction that just took place and the cellular resources involved.

This algorithm allows studying the dynamics of the simultaneous translation of different mRNA species assuming a fixed total amount of tRNA and ribosomal resources

($tRNA^T$ and R^T). The algorithm accounts for (i) the need of ribosome binding space on the beginning of the decoding region for translation initiation to take place, for (ii) the *traffic jam effect* due to ribosome queuing when slower codons are being translated, and for (iii) the fluctuations between active and free ribosomes (R^a and R^f).

2.2.1.3 Extension of stochastic framework to include tRNA abundance fluctuations

We improved the stochastic algorithm from (76) to account for fluctuations between active and free tRNA molecule abundances ($tRNA^a$ and $tRNA^f$) by allowing the dynamic tracking of the position of the tRNA molecules inside the ribosome during translation. We tracked the tRNA molecules bound to each ribosome translating a particular mRNA sequence between the following kinetic states (see Figure 2.1 in the section 2.2.1.1):

- From state 2 until A-site tRNA release at state 1 after initial binding (A-site OFF);
- From state 2 until A-site tRNA release at state 1 during proofreading following state 5 (A-site PROOF);
- From state 2 until P-site tRNA accommodation at state 10 (P-site ON), which corresponds to the time that it takes for a tRNA to complete a translation cycle of a codon, leaving the A-site free for the next codon to be translated;
- From state 2 until E-site tRNA release at state 1 following state 11 (E-site OFF), which corresponds to the passage of a tRNA from initial A-site binding, to the P-site after translocation, A-site binding of another tRNA species, translocation of the following codon and the newly bound tRNA species to P-site and consequent transfer of the previous tRNA and decoded codon to the E-site from where it is finally released. Note that during E-site OFF two P-site ON events have occurred for two consecutive codons.

A list with the tRNA-codon binding interactions is presented in Tables C.2.2 and C.2.3. From the possible tRNA choices, we selected the species that will participate in the binding reaction based on a distribution that takes into account the number of available molecules for each species at the time of the binding. Once a tRNA species is selected for binding with the mRNA-ribosome complex its amount decreases by one unit. Rejection

of the tRNA molecule during initial selection stage (states 1-3 or A-site OFF) and proofreading stage (states 5-1 or A-site PROOF), or simple deacylated-tRNA release from the ribosomal E-site (states 11-1 or E-site OFF) will result in the increase of the respective tRNA species amount by one unit.

2.2.2 Translational resources and mRNA cell composition

The concentrations of each tRNA species in *E. coli* at growth rates 0.4, 0.7, 1.07, and 1.6 h^{-1} were obtained from the experiments reported in (59). The tRNA concentration per codon and binding interaction is summarized in Table C.2.4 for the growth rate of 1.07 h^{-1}. We estimated the total number of tRNA molecules ($tRNA^T$) and ribosomes (R^T) per cell at each growth rate from the values reported in (58). The $tRNA^T$ was used with the total tRNA concentration ($[tRNA^T]$) to compute the respective cell volumes at the given growth rates (Eq. A.2.1). The values obtained for the cell volumes were inside the range determined for *E. coli* in (77) (see section A.2.1 and Figure B.2.1 for further details on this estimation and Table C.2.5 for values).

Finally, the number of tRNA molecules for each of the tRNA species at each growth rate was computed from their respective concentrations and the cell volume. We computed the average number of mRNA copies per *E. coli* cell at each growth rate from the mRNA synthesis rate per cell in function of growth rate reported in (58) (see section A.2.2 and Figure B.2.2 for further details on this estimation and Table C.2.5 for values).

Since we lack data on the mRNA sequences and respective copy numbers expressed at each of the growth rates under study (for which we have available tRNA concentration data), we constructed the mRNA pools of the cell at each condition by formulating a homogeneity criterion based on the fact that *E. coli* expresses mRNA in low copy number (78). This criterion assumes that the mRNA pools are qualitatively similar across the four growth rates and enforces them to approximate both the average mRNA length and the codon usage frequency (CU) of *E. coli*. The CU_j is a measure of the fraction of each codon j present in the genome of an organism, and thus independent of growth rate. The validity of this assumption was shown by comparing the mRNA expression in *E. coli* at low (79) and high (80) growth rates (Table C.2.6). The complete formulation of the homogeneity criterion is explained in section A.2.2. Briefly, from the

list of *E. coli* mRNA species encoding proteins only and excluding pseudogenes obtained from the *E. coli K12* strain in EcoGene 3.0 (81), we selected, based on the criterion, a subset of the listed mRNA species with 52% of the sequences in this subset classified as essential genes. Although, the identity of the mRNA species was preserved in the cell at the different four growth rates, the individual copy numbers were varied to match the estimated average number of mRNA sequences per cell. The *E. coli K12* CU was obtained from the Genomic tRNA database (82). In order to differentiate between the CU based on the genome and the codon usage frequency based purely on the set of mRNA copies present in a cell, we defined the *mRNA codon usage frequency* (mCU) as a measure of the fraction of each codon j (mCU_j) present in the mRNA pools at each growth rate, and whose values we enforced to approximate the ones of *E. coli* CU (see Figure B.2.3 for a comparison between CU and mCU at each growth rate).

The concept of *interaction-based mRNA codon usage frequency* ($IBmCU_{tRNA_i}$) is introduced here to quantify the frequency of codons in the system that interact with tRNA species i and that classify under a certain base-pair binding interaction. The $IBmCU_{tRNA_i}^{bi}$ is computed with the following expression

$$IBmCU_{tRNA_i}^{bi} = \sum_{codon\ j\ with\ bi\ for\ tRNA_i} mCU_j, \tag{2.1}$$

where mCU_j is summed over all codon species j that bind to the tRNA species i with the binding interaction bi. As mentioned above, there are four possible binding interactions (*WC*, *WB*, *nc*, *non*) and we further defined a fifth one to account for all cognate binding interactions: $IBmCU_{tRNA_i}^{cogn} = IBmCU_{tRNA_i}^{WC} + IBmCU_{tRNA_i}^{WB}$.

2.2.3 Codon elongation rate

We derived an expression for the *codon elongation rate* (k_{eff}) in function of the free cognate (*WC* and *WB*), near-cognate (*nc*) and non-cognate (*non*) tRNA concentrations and the ribosome kinetic parameters. This derivation was based on a deterministic model of translation (51), which was extended to account for the differentiation between two types of cognate binding interactions, for the possibility of *nc* mis-incorporation and for tRNA rejection at the proofreading stage (see section A.2.3).

Inserting the values of the kinetic rate constants from TableS1 we obtained an expression to compute k_{eff}^{j} for each codon *j*

$$k_{eff}^{j}[s^{-1}] = \frac{T_{WC,j}^{f}+0.5884\cdot T_{WB,j}^{f}+2.6233\cdot 10^{-4}\cdot T_{nc,j}^{f}}{0.0104[\mu M\cdot s]+0.4556[s]\cdot T_{WC,j}^{f}+0.0613[s]\cdot T_{nc,j}^{f}+0.0171[s]\cdot T_{non,j}^{f}}, \qquad (2.2)$$

where the variables are the free *WC*, *WB*, near-cognate, and non-cognate tRNA concentrations to codon *j*.

2.2.4 Simulation of translation in *E. coli* cell

We simulated the translation dynamics for an *E. coli* cell using the cell composition at 37℃ for the following growth rates 0.4, 0.7, 1.07, and 1.6 h^{-1} and the ribosome kinetics described in section 2.2.1.1. The termination rate constants (k_T) values were kept high for all mRNA species so that it did not constitute a rate-limiting step in translation, as observed in (83, 84). The translation initiation rate constants (k_I) for each mRNA species where calibrated such that the system reached a 80% ribosome activity in each simulated pool as estimated in (58) (see section 2.2.4.1).

In parallel, we simulated the heterologous protein expression of seven synonymous Firefly Luciferase transcripts introduced in an *E. coli* cell growing at 1.07 h^{-1}. Simulations were performed individually for each transcript and only one copy of the transcript was added to the pool of mRNA copies in a cell at 1.07h^{-1}. The sequence design of the transcripts was based on synonymous codon substitution that yields the same Luciferase amino acid sequence (see section 2.2.4.2 for further details on the mRNA sequences).

The data was extracted from the simulations during a time interval for which the system was at steady state (see example for 1.07 h^{-1}in Figure B.2.4 and Figure B.2.5). All simulation results were averaged over a large number of repetitions of the same simulation.

2.2.4.1 Calibration of the translation system to match literature parameters

Although genome-wide ribosome profiling data for *E. coli* has recently started to appear in the literature (80, 85, 86) that could be used to derive initiation rate constants (k_I) for

each mRNA sequence being expressed, these datasets do not exist for all the growth rates in this study along with its respective mRNA sequencing data. We have thus randomly attributed a k_I to the different mRNA species and subsequently multiplied them by a calibrating constant that differed across the different growth rates in order to reach 80% of ribosome activity (ribosomes that are being used for translation events) in each simulated pool as estimated in (58). Note that the use of this calibration constant leads to a difference in the initiation rate constants of the mRNA species across the four growth rates, which can, for instance, be taken as the result of changes in the amount of initiation factors. However, if the translation initiation efficiency of an mRNA species is higher than another, it will remain as such for all the growth rates as the initiation rate constants of each mRNA species were not changed individually. We remark that the calibration of the initiation rates to match the 80% level of ribosomes active in translation will always lead to a steady state with the same number of free ribosomes for each condition (if total ribosome amounts are fixed), which is independent of the individual ribosome profiles of each sequence and thus independent of the species of mRNA sequences present in the pool and their relative levels. High values for the termination rate constants (k_T) for all mRNA species were chosen in order not to limit the synthesis rate, as computational and experimental studies have shown for different organisms that translation of most mRNAs are initiation or elongation limited given experimental measurements of their polysome sizes (i.e. the number of ribosomes simultaneously translating an mRNA) (83, 84).

2.2.4.2 Heterologous expression of different Luciferase transcripts

We simulated the heterologous translation of seven synonymous Firefly Luciferase transcripts (one of them a wild type sequence) in an *E. coli* cell with growth rate 1.07 h^{-1}. Simulations were performed individually for each transcript and only one copy of the transcript was added to the pool of mRNA copies in a cell at 1.07h^{-1}. For each transcript, we fixed the termination rate constant (k_T) to a high value and the initiation rate constant (k_I) to the average k_I from all mRNA species used. We used six criteria for the design of the mRNA sequences based on synonymous codon substitution that yield the same Luciferase amino acid sequence:

- *WC & tRNA genes*: we replaced the codons in the Wild Type sequence with existing synonymous ones that are decoded by WC interactions and at the same time have the highest number of cognate WC tRNA genes. In case of tie in the number of genes for multiple synonymous codons, the codon with the highest number of *WB* interactions was chosen.
- *CU based*: we replaced the codons in the Wild Type sequence with existing synonymous ones presenting the highest *E. coli* codon usage frequency (CU).
- *WB based*: we replaced the codons in the Wild Type sequence with existing synonymous ones that are translated only by WB decoding tRNAs and that have no WC decoding isoacceptor tRNAs. If more than one possibility existed we choose the one with less WB decoding tRNA isocacceptors.
- *WC based*: we replaced the codons in the Wild Type sequence that have only WB interactions with existing synonymous ones that are also decoded through WC interactions. When more than one possibility exists, we chose the one with lowest cognate WC tRNA concentration.
- *TC based*: we replaced the codons in the Wild Type sequence with existing synonymous ones that have the highest total cognate tRNA concentration. The total cognate concentration is the sum of the concentrations of all tRNAs cognate to the codon and independent of binding interaction (*WC*, *WB*).
- k_{eff}^{max} *based*: we replaced the codons in the Wild Type sequence with existing synonymous ones that had the highest codon elongation rate as computed using Eq. 2.2 A variation of this transcript was constructed where the first 20 codons were maintained equal to the WT ($k_{eff}^{20,max}$*based*).

The first three criteria are the same as proposed in (60) and we used the synonymous mRNA sequences and Wild Type Luciferase reported therein (see Table C.2.7 for the list of transcripts and complete sequences used). We simulated 4000 times the heterologous translation of each transcript in an *E. coli* cell at 1.07h^{-1}.

2.2.5 Analysis methods for the stochastic system and parameter definitions

We simulated the simultaneous translation of different mRNA species from *E. coli* at different growth rate conditions (0.4, 0.7, 1.07, and 1.6 h^{-1}) using the stochastic framework and parameters described above. The data to characterize the translation system was extracted from the simulations during a time interval for which the system was at steady state (see example for 1.07 h^{-1} in Figure B.2.4 and Figure B.2.5). The system was assumed to be at steady state when convergence over the simulation time was reached for: the protein synthesis rate ($V_p(\tau_{sim})$) from all mRNA species, the number of free ribosomes ($R^f(\tau_{sim})$) and the number of free tRNA molecules ($tRNA_i^f(\tau_{sim})$) of each species i. All simulation results were averaged over 100 repetitions of the same condition.

The *protein synthesis rate at steady state* (V_p^k) was obtained by performing a time-average of the number of proteins produced from the total amount of an mRNA species k over the steady state time interval defined above. Dividing V_p^k by CN_{mRNA}^k (i.e., number of copies of an mRNA species k in the cell) we obtained the *specific protein synthesis rate at steady state* (V_s^k) for each mRNA species k, which corresponds to the protein synthesis rate per mRNA copies of mRNA species k.

The *elongation rate at steady state* (v_r^k), which is the average codon elongation rate of an mRNA species k per ribosome translating it, was computed with the following expression

$$v_r^k = \frac{V_p^k \cdot L_{mRNA}^k}{P^k \cdot CN_{mRNA}^k}, \tag{2.3}$$

where P^k is the polysome size of mRNA species k and L_{mRNA}^k is the length of the mRNA species k given by the number of codons between its start and stop codons.

The *ribosomal density* (ρ) is a measure of the fractional ribosome occupancy along an mRNA strand, expressed between 0 and 1. We defined it as in Eq. 1.4 with the following expression

$$\rho^k = \frac{P^k \cdot L_R}{L^k_{mRNA}}, \tag{2.4}$$

where L_R is the length of the 70S ribosome complex in terms of number of mRNA codons it occupies, which is assumed to be about 12 codons (87-89). We compute ρ^k for each mRNA species k as the time-average of its ribosomal density over the steady state time interval defined above.

The *mean distance between ribosomes* (D_R) is the average number of nucleotides that lay in-between the back and the head of consecutive ribosomes, assuming that the ribosomes are equally spaced along each mRNA sequence at steady state. We computed it with the following expression

$$D_R = 3 \cdot \frac{\sum_{k=1}^{all\ species}(L^k_{mRNA} \cdot CN^k_{mRNA}) - \sum_{k=1}^{all\ species}(P^k \cdot CN^k_{mRNA} \cdot L_R)}{\sum_{k=1}^{all\ species}(P^k \cdot CN^k_{mRNA})}, \tag{2.5}$$

where the multiplying factor 3 converts number of codons into number of nucleotides.

The steady state R^f and $tRNA^f$ for each species i were obtained by performing a time-average over the steady state time interval. We defined the *tRNA activity* as the steady state percentage of the total number of molecules of each tRNA species i that is active in translation events and consequently not available for translation, which is given by the following expression

$$tRNA_i\ activity = \frac{tRNA_i^a}{tRNA_i^T} \times 100, \tag{2.6}$$

where the number of active tRNA molecules at steady state is given by $tRNA_i^a = tRNA_i^T - tRNA_i^f$.

Translation time profiles inform about the time a ribosome spends translating each codon along an mRNA sequence. The time at which a ribosome starts translating each codon was recorded during the steady state interval defined above for each ribosome that translated an mRNA copy in our simulated cell. These were subsequently averaged to generate a translation time profile for each mRNA species separately. The

translation time profiles of each mRNA species in the system were broken down and the time intervals for the translation of each codon were grouped by codon species. The *codon elongation rate* obtained from our stochastic simulations (k_{stoch}^{j}) of each codon species *j* was then computed by averaging all the times spent by a ribosome to translate codon *j* and finding its reciprocal.

2.2.6 *In-silico* pulse-chase

We performed *in-silico pulse-chase* during translation of each of the seven transcripts by post-processing the translation time profiles (see section 2.2.5 for the definition of translation time profiles) of each of the seven Luciferase synonymous transcripts. The principle of *in-silico pulse-chase* is very similar to the experimental pulse-chase analysis. Using the translation time profiles, we counted, at each time point that a ribosome finished translating the complete mRNA sequence, the number of methionine amino acids that were incorporated in each complete Luciferase protein and whose methionine codons were translated during a fixed "labeling time", which in the experiments corresponds to the time for which the system has labeled methionine. The ribosome kinetics used in our system was measured *in vitro* at 20-25°C, whereas *in vivo* experiments take place at 37°C. Since the elongation rate depends on the temperature (58), it is necessary to calibrate our "labeling time" with respect to the typical experimental labeling time of 10s (90). During this 10s, approximately 100 codons are translated by a ribosome on a transcript presenting an elongation rate of 10 aa/s (value estimated in (60) for the translation of Wild Type Luciferase in an *E. coli* cell). In order for a ribosome in our system to translate 100 codons it was required a "labeling time" of 232s. The methionine level was normalized by the ratio between the number of methionines present in the protein sequence (14 in total) and the maximum level of methionine observed in the experiment. For comparison between the *in-silico* pulse-chase and the experiments, the time axis was multiplied by a factor of 23s representing the ratio between the time that takes for the translation of 100 codons of WT Luciferase in our simulation and in the experiments.

2.2.7 Time lags of ribosome occupancy by the tRNAs

The *ribosome occupancy time lag* ($\Delta t_{i,j}^{ds,bi}$) represents the time duration for a tRNA species i to reach one of the four decoding stages (ds) when bound to a ribosome on codon species j, with which it forms a specific base-pair binding interaction (bi) from the four possible ones (*WC*, *WB*, *nc*, *non*). The decoding stages are A-site OFF, A-site PROOF, P-site ON and E-site OFF as indicated in Figure 2.1 of the section 2.2.1.1. The ribosome occupancy time lags and their respective *number of events* ($n_{i,j}^{ds,bi}$) were recorded during the successive translations along an mRNA sequence. We computed the *mean ribosome occupancy time lag per decoding stage and binding interaction* with

$$\overline{\Delta t}_{bi}^{ds} = \frac{\sum_j \sum_i \left(\Delta t_{i,j}^{ds,bi} \cdot n_{i,j}^{ds,bi}\right)}{N_{bi}^{ds}}, \tag{2.7}$$

where $N_{bi}^{ds} = \sum_j \sum_i n_{i,j}^{ds,bi}$ is the *total number of events per decoding stage and binding interaction*.

We estimated the *average time of codon translation per incorporated amino acid and binding interaction* (t_{codon}^{bi}) by dividing the total decoding time per binding interaction by the number of amino acids that were successfully incorporated in the protein sequence during multiple translations, which is represented by the total number of E-site OFF (or equivalently P-site ON) events

$$t_{codon}^{bi} = \frac{\sum_{ds} \overline{\Delta t}_{bi}^{ds}}{N_{WC}^{E-OFF} + N_{WB}^{E-OFF} + N_{nc}^{E-OFF}}. \tag{2.8}$$

2.3 Results and Discussion

2.3.1 General translation properties of the cell in function of growth rate

We studied the distribution of the *protein synthesis rate* (V_p) for each growth rate (Figure 2.2a). The mean V_p among the mRNA species is observed to increase with the growth rate, along with the increase in translation resources (see Table C.2.5). Interestingly, the mean *elongation rate* (υ_r) (Eq. 2.3 in section 2.2.5) is observed first to increase and then

decrease with growth rate, which is accompanied by a respective decrease and increase in the mean *ribosomal density* (ρ) (Eq. 2.4 in section 2.2.5), contrary to the estimations in (58) where it increases with growth rate (Figure 2b). Despite this difference, we found good agreement between the *mean distance between ribosomes* estimated in (58) and our computed values (Eq. 2.5 in section 2.2.5) for each growth rate using the ρ values of each mRNA species from our simulations and assuming the ribosome has the size of one nucleotide $L_R = 0$ (Figure B.2.6a). However, the real length of the ribosome covers about 12 codons, which is accounted for in our simulations, and performing the computation with the correct ribosome length shows that the values from the literature overestimate the true mean distance between ribosomes. The effect observed above results directly from the inverse changes in mean ρ given that the mean *tRNA activity*, i.e., the fraction of tRNA species bound to ribosomes (Eq. 2.6 in section 2.2.5), of all tRNA species is similar for the different growth rates (Figure B.2.6c) and the average of the fold changes on abundance from all tRNA species at each growth rate relative to $0.4h^{-1}$ is fairly constant (Figure B.2.6d) and suggests no major changes in the ratio between cognate and competitor tRNA concentrations that could affect υ_r.

Interestingly, with the increase in growth rate, the V_p distribution shifts from having one peak to a bimodal distribution and back again to having just one peak, suggesting a systemic change in the control of the synthesis rate. Previous computational and experimental studies have shown for different organisms that translation of most mRNA species is initiation or elongation limited (83, 84). In particular, computational studies have shown that the *specific protein synthesis rate* (V_s), i.e., the rate of proteins produced per number of copies of an mRNA species (see section 2.2.5), is limited by translation initiation for low values of ρ, by translation termination for high values of ρ, and reaches a maximum for moderate values of ρ for which translation elongation becomes limiting (42, 51, 57). We observe that the cells simulated at growth rates $0.4h^{-1}$ and $1.6\ h^{-1}$ have a higher number of mRNA species that cluster in a more stable region of the V_s curve with higher ρ and higher initiation rate ($k_I \cdot R^f$) indicating that these are mostly elongation limited, whereas the cells at growth rates $0.7h^{-1}$ and $1.07h^{-1}$ have their mRNA species clustered into two groups that correlate with the V_p bimodal distributions: one group with lower ρ for which translation is mostly initiation limited and a second group with higher ρ values for which translation is mostly elongation limited (Figure 2.2c-d). A

decrease bigger than 20% in the ratio between R^f and the total number of mRNA copies at 0.7h^{-1} and 1.07h^{-1} with respect to the one at 0.4h^{-1} suggests a limitation in free ribosome resources (Figure B.2.6b). Thus, it appears that under low and high growth rates the system optimizes protein translation with higher V_s for the mRNA species, whereas for intermediate growth rates, translation initiation regulates protein synthesis. Consequently, at growth rates 0.4h^{-1} and 1.6h^{-1} there is a higher proportion of mRNA species limited by translation elongation, limitation that corresponds to more ribosome blocking (higher ρ) due to queuing of ribosomes downstream the sequences and lower v_r, so that the overall mean v_r of the entire mRNA pool decreases (Figure 2.2b).

Although these results were obtained from simulations considering a homogeneous mRNA pool across mRNA conditions (see section A.2.2), they are valid for any choice on pool composition as long as the total number of ribosomes and mRNAs in the system remains as parameterized here. This is consequence of the calibration performed on the initiation rates to force the cell at each growth rate to have 80% of its ribosomes active in translation. This leads to a steady state where for each growth condition the level of free ribosomes will always consist of the remaining 20%, which is independent of the individual ribosome profiles of each sequence. If the ratio between free ribosomes and total number of mRNAs is maintained independently of the mRNA pools used, the shifts on average ribosomal densities will be observed since they constitute a direct effect of the competition among translating mRNAs for free ribosomes, which directly influences the level of ribosome crowding along the mRNA sequences.

2.3.2 Determinants of elongation rate

We investigate further the determinants of translation elongation rate by focusing on the analysis of the cell at growth rate 1.07 h^{-1} (highest mean v_r and moderate mean ρ) and the production of a heterologous protein. The results discussed below are similar for the four growth rates and thus independent of the changes in ρ, unless otherwise stated.

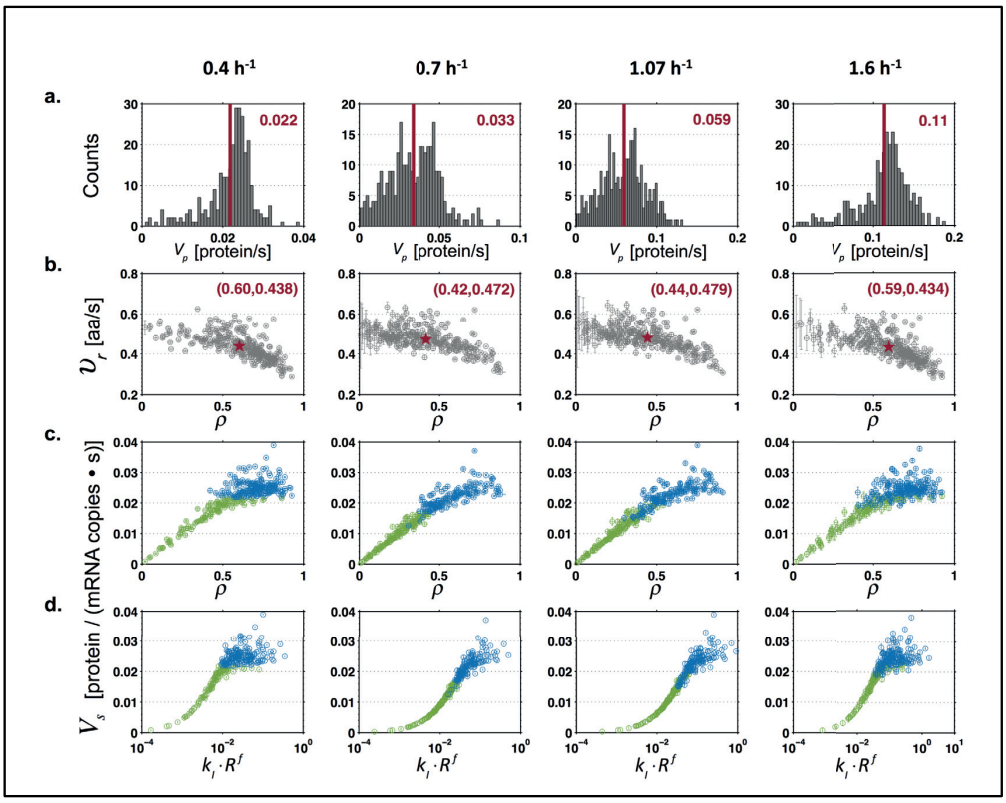


Figure 2.2 (a) Distribution of protein synthesis rate (V_p) for the different growth rates. Red bar and number represent the mean V_p among all mRNA species. (b) Elongation rate (υ_r) for each mRNA species in function of the ribosomal density (ρ) for the different growth rates. The red star and text represent the mean (ρ, υ_r) from all the mRNA species in the cell. Vertical and horizontal error bars represent standard deviation from 100 repeated simulations. (c-d) Specific protein synthesis rate (V_s) for each mRNA species in function of the ribosomal density (ρ) (c) and in function of the initiation rate ($k_I \cdot R^f$) (d) for the different growth rates. Green and blue color code separates the data points that have a V_p below or above the mean V_p among all mRNA species, respectively. Vertical and horizontal error bars represent standard deviation from 100 repeated simulations.

In order to qualitatively validate the model and its parameters, we separately simulate the translation of four Luciferase transcripts in an *E. coli* cell growing at 1.07 h^{-1}, we post-process the translation time profiles of the transcripts (see sections 2.2.5 and 2.2.6) and we compare our results with the ones from pulse-chase experiments performed by Spencer and colleagues (60). For the simulations we use the same Luciferase transcripts as in (60), which consist of a wild type (WT) Luciferase transcript and three other sequences whose designs follow different criteria based on synonymous codon substitution: codons with existing WC decoding tRNA isoacceptors combined

with high tRNA gene copy number (*WC & tRNA genes*), codons with the highest genome codon usage frequency (*CU based*), and codons without *WC* decoding tRNA isoacceptors (*WB based*) (criteria detail in section 2.2.4.2 and mRNA sequences in Table C.2.7). The mean time-evolution curves of methionine level from our *in-silico pulse-chase* performed on the WT, *WC & tRNA genes* and *CU based* transcripts (no experimental curve available for *WB based*) show good agreement with the experimental curves from (60) (Figure 2.3a). The experimental curves obtained at 37°C and hence with faster elongation rate were calibrated for comparison with our system at 20-25°C by multiplying the time axis by a factor of 23s (see section 2.2.6). The deviations between the simulated (sim) and experimental curves are accounted by the distribution of the simulated curves that generated our mean time-evolution curves of methionine level.

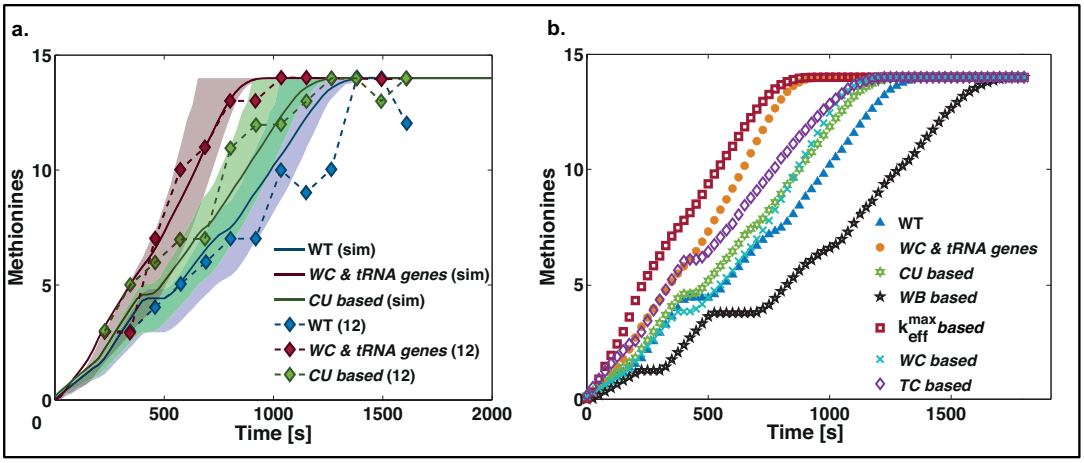


Figure 2.3 (a) Comparison between simulated (sim) and experimental time-evolution curves of methionine level obtained from experiments in (60) for WT Luciferase and for two of its synonymous transcripts (*WC & tRNA genes* and *CU based*). Bounds represent the 25th and 75th quartile of the distribution from the *in-silico pulse-chase* curves. Time axis from the experimental data points was adjusted with the same calibration factor used for the methionine "labeling time" (see section 2.2.6). (b) *In-silico pulse–chase* performed during the translation of seven heterologous transcripts yielding the same amino acid sequence based on different synonymous codon substitution criteria in *E. coli* cells at 1.07 h^{-1}. The time-evolution curves of methionine level result from the average of 4000 repeated simulations. The curves are plotted with the bounds representing the 25th and 75th quartile of their sample distribution in Figure B.2.7.

Three other Luciferase transcripts were designed where codons were replaced by their synonyms based on existing *WC* decoding tRNA isoacceptors (*WC based*), based

on the highest cognate tRNA concentration (*TC based*), and based on the highest codon elongation rate (k_{eff}^{max} *based*). For the design of the k_{eff}^{max} *based* transcript we computed the *codon elongation rate* (k_{eff}^{j}) of each codon species j with Eq. 2.2, using the steady state tRNA concentrations obtained after simulating an *E. coli* cell at 1.07 h^{-1}. The *WC & tRNA genes* and k_{eff}^{max} *based* transcripts present the fastest elongation rates (Figure 2.3b; for translation time profiles see Figure B.2.8) and highest protein synthesis rates when compared to the WT transcript (28% and 28.6%, respectively, see Figure B.2.9a), with k_{eff}^{max} *based* being more optimal. These two transcripts differ only in the use of two codon species (one encoding for Glutamine and the other for Serine) that, combined, appear in 43 positions along the 585-codon sequence. The similarity between these two transcripts is explained by the previously observed correlation between tRNA abundance and its gene copy number (59, 91, 92), and the fact that k_{eff} of a codon is maximized by high concentration of cognate *WC* tRNA and low competition. Only the *WB based* transcript leads to a decrease in protein synthesis (about 20% less translated protein, see Figure B.2.9a) relative to the WT transcript. We tested a transcript where the 20 first codons were not changed and confirmed that the different pulse-chase curves between WT and k_{eff}^{max} *based* are a result of changes on elongation rate rather than initiation (Figure B.2.9b). Nevertheless, we note here that even though a transcript is optimized for elongation by synonymous codon substitution with the purpose of increasing protein production levels, the translation initiation rate, which is dictated by the beginning of the transcript's sequence and the steady state R^f of the host cell, has a major impact on the gain in protein production with respect to the WT in its rate limiting regime (as seen above with the specific protein synthesis rate (V_s)) and is further discussed in Figure B.2.9b-d.

These findings are supported by sensitivity analysis of the ribosome kinetic parameters with respect to υ_r. After performing an initial screening on the 25 kinetic rate constants to identify the insignificant ones (Figure B.2.10), we use a Monte-Carlo based numerical procedure for variance-based global sensitivity analysis (93) and determine the ribosome kinetic rate constants that influence the most the value of υ_r observed for an mRNA transcript (Figure 2.4a) (results are valid for any mRNA species as ribosomal kinetic pathway is the same for all codons). The analysis shows no

dominant rate constants (all sensitivity indices < 0.4), but their order of influence ranked by their main effects (S_{k_i}) indicates that k_{-1}, k_{-2}^{nc} and k_5^{WB} are the most influential rate constants, which indicates that there are two decoding stages of the ribosome that determine v_r: (i) rejection of competitor tRNA (k_{-1} and k_{-2}^{nc}) during initial selection, and (ii) accommodation of cognate WB tRNAs (k_5^{WB}). Interestingly, although the influence of k_2 on v_r is mostly due to interaction effects ($S_{T_{k_i}}$), we observe that the nominal value (obtained from experiments) of k_2 seems to be optimized to yield the highest v_r (inset in Figure 2.4a). Analysis of the system for the parameters at 37°C and for the *in vivo* parameters deduced from *in vitro* ones (94), showed a redistribution of the rate limiting steps (Figure B.2.11a-b), where cognate binding interaction (k_5^{WB}) becomes more important than the overall tRNA competition (k_{-1} and k_{-2}^{nc}). This finding supports the discussions on tRNA competition as observed at *in vitro* conditions presenting an inhibitory effect on translation elongation that would decrease translation efficiency if maintained at *in vivo* conditions (94, 95). The fact that the ratio between each ribosome kinetic rate constant at 37°C and at 20°C is approximately the same (except for the initial tRNA binding rate constant and at least until the tRNA accommodation for which we have values to compare) explains why our results at 20-25°C match so well the experiments performed at 37°C (60). We note here that when we combine the ribosome kinetics at 20-25°C with the parameters of the system at 37°C (such as total number of ribosomes, total number of tRNAs and total number of mRNAs) we also perform a scaling of the initiation rate to bring the translation process to the conditions at 37°C by enforcing 80% ribosomes to be active in translation. If in this system, the ribosome kinetics were to be replaced by a faster kinetics, such as the one at 37°C, the resulting effect would be a faster elongation rate due to the ribosomes faster movement along the sequence and a faster update of the number of free ribosomes. This is so because elongation rate at 37°C is expected to increase, and such increase would happen uniformly for all sequences as the ribosome kinetics is assumed to be equal for all codons. However, if the system was to be submitted to the same scaling condition on initiation that enforces an 80% ribosome activity, the steady state reached for this system would be the same as the one obtained with a ribosome kinetics at 20°C.

These results support the optimality of k_{eff}^{max} *based* transcript design as less competition and higher rate of accommodation for cognate *WB* improves k_{eff} values. Furthermore, the high correlation between the *codon elongation rates* obtained from our stochastic simulations (k_{stoch}) (see translation time profiles definition in section 2.2.5) and from the deterministic model (k_{eff}) (Figure 2.4b) indicates that the latter is a high accuracy predictor of codon elongation rate of slow codons (i.e., codons limiting translation). Stochastic queuing effects that are dependent on the mRNA sequence downstream introduce variability on the measured k_{stoch} for fast codons and bias the codon elongation rate towards values that are lower than the ones expected in a theoretical system without ribosome queuing interference.

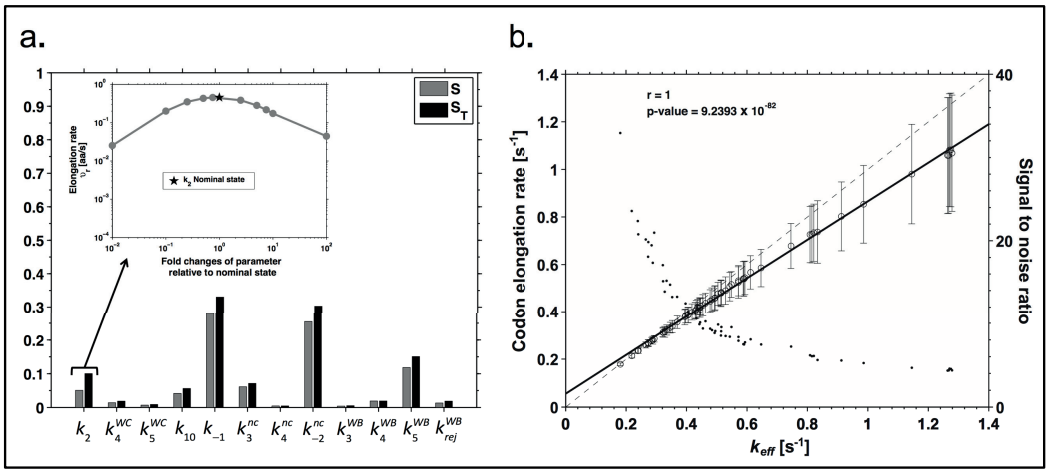


Figure 2.4 (a) Main and total effects (S_{k_i}, $S_{T_{k_i}}$) on the value of v_r due to a change in ribosome rate constant k_i. Inset: Changes in v_r in function of the changes on k_2 for a range of two orders of magnitude below and above its nominal value (star). (b) Codon elongation rate obtained from stochastic simulations (k_{stoch}) *vs.* the codon elongation rate constant (k_{eff}) (open circles). Each data point corresponds to one of the 61 codons taking part in the translation elongation. The linear regression line is represented by continuous line. The Pearson correlation coefficient (r) and p-value are indicated. The dashed line is the one-to-one function for comparison. The signal to noise ratio from k_{stoch} corresponding to each codon is represented by the dots and remains higher than 1 for all codons, starting to stabilize for the codons with higher elongation rates.

2.3.3 Key factors on tRNA activity

The amount of tRNA available for translation used to estimate k_{eff} dictates both the cognate and competitor tRNA concentrations for each codon, and directly depends on

the amount of tRNA that is active in translation, i.e., occupying the ribosomes. We estimate the *mean ribosome occupancy time lag* ($\overline{\Delta t}_{bi}^{ds}$) and the *total number of events* (N_{bi}^{ds}) *per decoding stage (ds) and binding interaction (bi)* using Eq. 2.7 for the WT Luciferase transcript in an *E. coli* cell growing at 1.07 h^{-1}. The decoding stages are A-site OFF, A-site PROOF, P-site ON, and E-site OFF (see Figure 2.1 and section 2.2.7), and the possible binding interactions are *WC*, *WB*, *nc*, and *non*. The statistics obtained here are valid for any mRNA sequence and growth rate. Higher ribosomal densities increase both $\overline{\Delta t}_{bi}^{P-ON}$ and $\overline{\Delta t}_{bi}^{E-OFF}$ because of slower translocation of the ribosome, but the proportions among the events remain the same (results not shown). Note that $\overline{\Delta t}_{bi}^{ds}$ values result directly from the intrinsic ribosome kinetics and, as such, they are very similar for all the different tRNA-codon interactions, except for P-site ON and E-site OFF decoding stages where the ribosome translocation time depends on the ribosome queuing downstream the mRNA sequence (Figure B.2.12), whereas the number of events depends on the codon species and the free tRNA abundances (Figure B.2.13). Most of the tRNAs involved in cognate binding interactions (68.74% and 40.35% for *WC* and *WB*, respectively) are accepted for the peptidyl bond formation and occupy the ribosome until its release at the E-site after a long $\overline{\Delta t}_{WC}^{E-OFF}$ or $\overline{\Delta t}_{WB}^{E-OFF}$ has occurred (Table 2.1). Thus, tRNA species with higher *cognate-based mRNA codon usage frequency* $IBmCU_{tRNA_i}^{cogn}$ (see Eq. 2.1) have also higher frequency of events that result in E-site release, and subsequently are active in translation in higher amounts as shown by the correlation found in Figure 2.5. However, there is a difference of about 28% between *WB* and *WC* binding interactions that will not reach decoding stage E-site OFF and will instead end up with the tRNA being rejected at proofreading stage (A-site PROOF), which is a much faster event than for E-site OFF. Deviation from the regression line corresponds to cases for which the proportion of $IBmCU_{tRNA_i}^{WC}$ is very low, and $IBmCU_{tRNA_i}^{WB}$ is not high enough to compensate for the number of tRNA molecules that could be active if there was high $IBmCU_{tRNA_i}^{WC}$ instead of $IBmCU_{tRNA_i}^{WB}$. Such are the cases of the outliers Leu3, Pro3, Val2A and Val2B in Figure 2.5 (see proportion of $IBmCU_{tRNA_i}^{WC}$ and $IBmCU_{tRNA_i}^{WB}$ in Figure B.2.14). Thr1 deviates from the regression line because it is the species with the lowest concentration in the cell and with an abundant isoacceptor (Thr3) (Figure B.2.15). Therefore, the probability of Thr1 to bind with the ribosome is

reduced and hence its activity is not representative of the $IBmCU_{Thr1}^{cogn}$. Because a small number of near-cognate binding interactions can reach A-site PROOF and E-site OFF decoding stages (1.08% and 0.02%, respectively), $IBmCU_{tRNA_i}^{nc}$ can be high enough such that the number of A-site PROOF or E-site OFF events can compensate for a low $IBmCU_{tRNA_i}^{cogn}$ or a high proportion of $IBmCU_{tRNA_i}^{WB}$with respect to $IBmCU_{tRNA_i}^{WC}$, and hence contribute to a higher tRNA activity (which is the case of Val1 in Figure B.2.14).

Table 2.1 Statistics on mean ribosome occupancy time lags and total number of events per decoding stage and binding type.

	WC	***WB***	***nc***	***non***
$\overline{\Delta t}_{bi}^{A-OFF}$ **[×10^{-2} s]**	0.36 (0.006)	0.48 (0.009)	3.9 (0.009)	1.2 (0.0008)
N_{bi}^{A-OFF}	70961	86665	4651313	15757894
% *	30.92	31.8	98.9	100
$\overline{\Delta t}_{bi}^{A-PROOF}$ **[s]**	0.16 (0.003)	0.51 (0.01)	0.31 (0.001)	-
$N_{bi}^{A-PROOF}$	790	75871	50877	-
% *	0.34	27.85	1.08	-
$\overline{\Delta t}_{bi}^{P-ON}$ **[s]**	0.70 (0.01)	1.00 (0.02)	0.83 (0.002)	-
$\overline{\Delta t}_{bi}^{E-OFF}$ **[s]**	3.09 (0.05)	3.41 (0.07)	3.29 (0.008)	-
$N_{bi}^{P-ON/E-OFF}$	157731	109924	843	-
% *	68.74	40.35	0.02	-
t_{codon}^{bi} **[s/aa]**	1.82	1.54	0.74	0.69

* Fraction of *bi* events (*WC*, *WB*, *nc*, *non*) per decoding stage (A-site OFF, A-site PROOF and P-site ON/E-site OFF).
(values) are standard deviations.

From N_{bi}^{ds} presented in Table 2.1 we compute a total of 2.4% of cognate binding events among all possible binding events, and only 1.3% of these 2.4% resulted in a complete codon translation (reaching P-site ON, thus leaving the A-site free for the binding of a new tRNA). The larger bulk of translation binding events consists of interactions with competitor tRNAs (97.6%), which supports tRNA competition as a determinant of elongation rate. We compute the *average time of codon translation per incorporated amino acid and per binding interaction* (t_{codon}^{bi}) using Eq. 2.8. The fact that t_{codon}^{non} is of the same order of magnitude as t_{codon}^{nc} (Table 2.1) implies that non-cognate binding interactions cannot be dismissed on the basis of these being fast events, contrary to the assumption made by Fluitt and colleagues (64), which is used in recent translation

modeling attempts (67, 68), where these are ignored based on their fast interactions and all competition in the system is resumed to near-cognate binding interactions. In fact, although the total time spent per non-cognate binding interaction is small ($\overline{\Delta t}_{non}^{A-OFF} \approx 1.2\times10^{-2}$s), one needs to take into account that these binding events have a very high frequency of occurrence. If we compute the ratio between the total time spent per near-cognate interaction and non-cognate interaction $\left(\frac{\overline{\Delta t}_{nc}^{A-OFF}\cdot N_{nc}^{A-OFF}+\overline{\Delta t}_{nc}^{A-PROOF}\cdot N_{nc}^{A-PROOF}+\overline{\Delta t}_{nc}^{E-OFF}\cdot N_{nc}^{E-OFF}}{\overline{\Delta t}_{non}^{A-OFF}\cdot N_{non}^{A-OFF}}\right)$ we observed that this value is very close to 1, implying that non-cognate events are as important as near-cognate ones. Nevertheless, despite the high number of near-cognate binding events, we estimate that near-cognate mis-incorporation occurs only once for each 318 cognate *WC* and *WB* complete codon translations, resulting in an error frequency of about 3×10^{-3}, which is in the range of the *E. coli in vivo* measurements (55, 56, 96-99). These results underline the significance of competition in the dynamics of translation elongation and they remain valid for *in vivo* conditions (Figure B.2.11) despite the observed decrease of the effect of tRNA competition in translation and subsequent decrease in error frequency of a factor of 3.

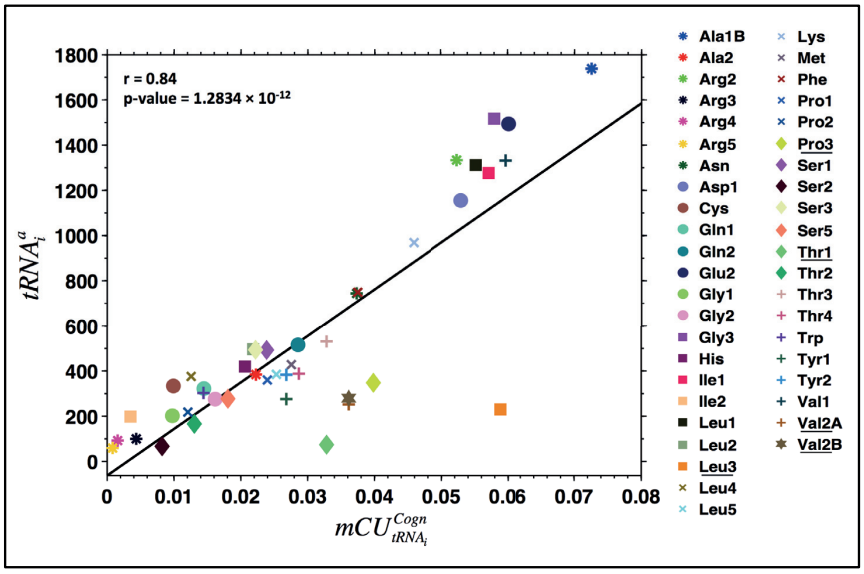


Figure 2.5 Number of tRNA molecules of species i active in translation ($tRNA_i^a$) in function of its respective cognate interaction-based mRNA codon usage frequency ($IBmCU_{tRNA_i}^{cogn}$). The Pearson correlation coefficient (r) and the p-value are indicated. Correlation outliers are underlined in the legend.

2.3.4 Global effects of amino acid starvation and surplus in the cell

Given the role that tRNA availability plays in elongation rate, an interesting question is how the surplus or starvation of certain amino acids will globally affect elongation rate in the cell. To answer this question, we simulate 20 times the cell at growth rate 1.07 h^{-1} and in each simulation, we increase or decrease the concentration of each tRNA isoacceptors for the same amino acid by 50% of their literature values at the given growth rate. Analysis of the relative deviation of the average elongation rate from all mRNA species with respect to the standard case at 1.07 h^{-1} shows three regimes according to the effect of the increase/decrease of the amino acid concentrations on the average elongation rate in the cell (Figure 2.6). Similar results were obtained when the concentrations where changed by 20% and 50% for each tRNA species separately (Figure B.2.16) and analysis of this results revealed the mechanisms behind the observed effects (details in Figure B.2.17). The amino acids Phe, His, Met, Asn, Pro, and Gln in regime (i) generally limit translation in the cell under starvation conditions and improve elongation rates under surplus. These amino acids have isoacceptor tRNAs that are among the ones whose cognate (specially *WB* type) codons have very slow codon elongation rates and present a low ratio between codon elongation rate and cognate codon mCU on the mRNA sequences in the cell (Figures B.2.18 and B.2.19). On the other hand, the amino acids Gly, Glu, and Arg in regime (ii) generally limit translation in the cell under surplus by increasing the competition on their near- and non-cognate codons, while improving translation under starvation conditions due to diminished competition pressure. These amino acids have tRNA isoacceptors that are among the species in the cell that are present in highest abundances (Figure B.2.15) and, as a consequence, their cognate codons have the highest codon elongation rates (Figure B.2.19). In the case of Leu, the negative effect on elongation rate due to the surplus of its tRNA isoacceptor Leu1 still prevails, however, the combined effect from all its isoacceptors under starvation is characteristic of group (iii). The amino acids in regime (iii) are the ones that have an effect similar to (ii) but with smaller extent under surplus due to the increase in competition resulting from their combined isoacceptor high abundances or high $IBmCU^{nc}_{tRNA_i} + IBmCU^{non}_{tRNA_i}$, however, under starvation their cognate codon elongation rates are negatively affected by the high $IBmCU^{cogn}_{tRNA_i}$ demanding free tRNA.

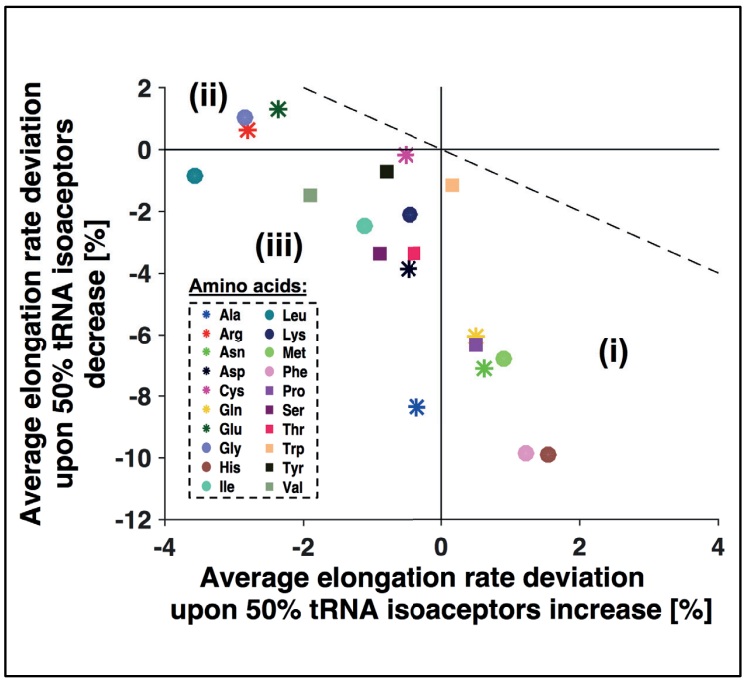


Figure 2.6 Each marker represents the relative change of mean elongation rate of the cell's transcriptome upon starvation (y-axis) or surplus (x-axis) of the respective amino acid with respect to the reference state (1.07h^{-1} growth rate). A simulation was performed for each amino acid surplus and starvation change to be compared to the reference state. Along the x-axis, the values represent the relative deviation of the average elongation rate from all mRNA species in the cell upon combined 50% increase of the abundance of all tRNA isoacceptors per amino acid type. Along the y-axis, the values represent the relative deviation of the average elongation rate from all mRNA species in the cell upon one-at-a-time 50% decrease of the abundance of all tRNA isoacceptors per amino acid type. One-to-one line (dashed) plotted for comparison.

Overall, starvation of a tRNA or an amino acid has a more pronounced effect on the cell's translation behavior because it acts upon the rate limiting codons. Nevertheless, the global effects of competition on translation elongation due to transient changes in nutrient supply introduce in the cell another level of regulation of the patterns of protein synthesis as a response to stress. Our stochastic framework has the potential to study the surplus/starvation effect of changes in the amount of tRNA competition on the elongation of the different mRNA species in detail. For instance, the most recent stochastic approach for modeling translation (68), is not able to observe the effect on mean elongation rate due to diminished competition pressure for group (ii) (amino acids Gly, Glu, and Arg), concluding instead that all types of amino acid starvation only lead to a decrease of the mean elongation rate in the cell.

2.4 Conclusion

In this work, we used a stochastic framework to model the translation process based on the available ribosome kinetics that describes tRNA competition while discriminating between cognate Watson-Crick and wobble interactions, and we simulated the simultaneous translation of a representative pool of *Escherichia coli* mRNA sequences under a range of different growth rates (0.4, 0.7, 1.07, and 1.6 h^{-1}) with parameters obtained from literature. The variation of the mean elongation rate from all mRNA species observed with the change in cell growth rate resulted from a systems response to an alteration of the ratio between free ribosomal resources and the number of mRNA copies that required those resources. Control of translation is observed to shift between initiation and elongation, which is characterized by a change in the ribosomal densities, and thus fine-tunes the protein synthesis of the mRNA species in the cell. We do not observe an increase in the mean elongation rate with growth rate as estimated in (58), from where the data was collected. The way mean elongation rate is usually estimated (58, 100) takes only into account the protein synthesis and the number of free ribosomes, which both increase with growth rate (not necessarily in a proportional way), and not the changes that can occur in ribosomal density and that affect elongation rate. Our results are consistent with what is expected from a system where *codon elongation rate* is determined by the ribosome kinetics and the free tRNA concentrations, and where *elongation rate* is determined by the combined effect of the multiple codon elongation rates and the ribosomal density along a sequence, which is also influenced by the initiation rate. For a system where tRNA competition effects are not observed to change radically with growth rate and the mRNA pool is qualitatively constant (such as here, despite the change in tRNA levels), the mRNA species being translated will present maximum elongation rate under initiation limiting conditions and lower elongation rate under elongation limiting conditions (if the ribosomal density is such that high ribosome queuing interaction impacts negatively elongation). These results suggest that the actual mean elongation rate is thus no longer well represented by just the amount of protein synthesis in the system, as for some mRNA species the highest protein synthesis is achieved by crowding the sequence with ribosomes, which may result in lower elongation rate. This implies that the actual elongation rate of some mRNA species may remain constant under different growth rates, whereas for others it

may decrease as a result of the level of ribosome crowding. This is consistent with the observation of an approximately constant elongation rate for *lacZ* for increasing growth rates (101).

Our sensitivity analysis results showed that the level of tRNA competition and the type of cognate binding interaction (*WC* vs. *WB*) determine elongation rate. The design of heterologous transcripts based on optimizing the sequence with synonymous codons that minimize tRNA competition and maximize the *WC* binding interactions with their cognate tRNAs was shown to lead to higher protein production. However, there is a trade-off between protein production level and elongation rate due to ribosome crowding effects. We proposed an equation to assist the design of optimized mRNA sequences that computes the codon elongation rate (k_{eff}) of a codon given that the amount of free tRNA species in the host organism is known. Nevertheless, since this equation will only help to design a sequence with codons that have high codon elongation rates, final protein specific activity will need to be tested, as it has been demonstrated that co-translational folding of proteins during the translation of slow codons is essential for correctly creating specific domains that determine the protein activity (26-28).

The analysis of our system showed that non-cognate binding interactions do in fact contribute to the competition level as much as the near-cognate ones do, contrary to the assumption made by Fluitt and colleagues (64) that these can be ignored based on their fast interactions and thus assuming that all competition in the system is resumed to near-cognate binding interactions. Furthermore, the existing stochastic models (67, 68) of translation use the latter results in order to estimate a factor for the tRNA competition, which is fixed per codon and is integrated in the codon elongation. This competition factor is estimated using the total amount for each tRNA species in the cell instead of the free transient tRNA amount that can be obtained at each step of the simulation, and as a consequence the effect of competition from changes in tRNA availability is no longer representative of the actual state of the cell. Because we accounted for these, we observed in our surplus/starvation studies the effect of changes in the amount of tRNA competition on the elongation of the different mRNA species.

Similar studies performed in (68) failed to observe this effect, concluding that all types of amino acid starvation only lead to a decrease of the mean elongation rate in the cell.

Furthermore, the results presented here, which were obtained from a ribosome kinetics at 20-25°C, were validated for higher temperature of 37°C, which is more consistent with *in vivo* conditions, and for the deduced *in vivo* kinetic rates obtained in (94). Analysis of the system for the parameters at 37°C and for the *in vivo* parameters deduced from *in vitro* ones (94), were found to support a translation model for which tRNA competition, although still an important factor, has a lower impact in translation elongation rate then the type of cognate decoding (94, 95).

In conclusion, our stochastic framework has proven to be effective in the analysis of a complex system such as translation. Literature describing the parameters for translation resources and specifically the ribosome kinetics is widely available for *E. coli*. The use of parameters and data pertaining to a specific organism establishes the framework for the study and modeling of systems with a number of components that correspond to the size of a biologically meaningful translation system, not needing to rely in the use of simplified parameter regimes. Furthermore, since the ribosome decoding center has been shown to be highly conserved among species during evolution (37) and similarities have been reported in the function of the different elongation factors in both bacteria and eukaryotes (102), the results observed in this work remain therefore valid for other organisms. This framework is a valuable tool for the systematic study of translation. Adding information on ribosome or polysome profiling experiments, as well as mRNA sequencing data for the specific conditions under study when available, can be valuable to the systems-level analysis of translation in the cell. This framework could thus be used for future work focused on: (i) exploring particular patterns of translation in certain mRNA sequences that could then be clustered by functionality and by codon frequency; (ii) studying the impact of changing the sequence of certain endogenous genes on the translation of the other mRNA sequences in the cell; (iii) studying the impact of expressing heterologous genes on the translation of other mRNA sequences.

Part II Establishing a pipeline for building reduced human metabolic models

Chapter 3: Pipeline for GEM processing and integration of data and thermodynamic parameters

3.1 Introduction

3.1.1 Genome Scale Metabolic Networks

GEnome-scale metabolic models (GEMs) have been extensively used in the field of Systems Biology for understanding the metabolic capabilities of organisms under different conditions (103-105). GEMs consist of a list of biochemical reactions representing the cellular metabolism of a given organism. These reactions form networks where the nodes are chemical compounds that can be either substrates, products or cofactors of said reactions. The rate at which these chemical transformations take place is represented by the flux of each reaction in the network.

The GEM of an organism is initially reconstructed by collecting all the available information on the metabolic genes of the organism and their respective gene-protein-reaction relationships (GPRs), which are annotated into the model structure and can be used for the association with transcriptomics/proteomics data. Gaps in the knowledge of the network are "solved" with further literature and/or gap-filling algorithms, whose purpose is to infer missing reactions by testing the ability of the network to perform cell/tissue specific metabolic tasks, i.e., ensuring that pathways for the production of certain compounds are active given the uptake of their precursors or main carbon sources. For the analysis, a GEM can be represented using a mathematical formulation where the reactions and respective participating metabolites are allocated into a matrix (the stoichiometric matrix) (Figure 3.1). This matrix has a column for each reaction in the network and a row for each metabolite. Each element in the matrix (the stoichiometric coefficient) represents the number of molecules of a given metabolite that are consumed (negative sign) or produced (positive sign) in each reaction. The zero elements in the matrix represent metabolites and reactions that are not connected. The metabolic reactions represented in the stoichiometric matrix are mass and energy balanced.

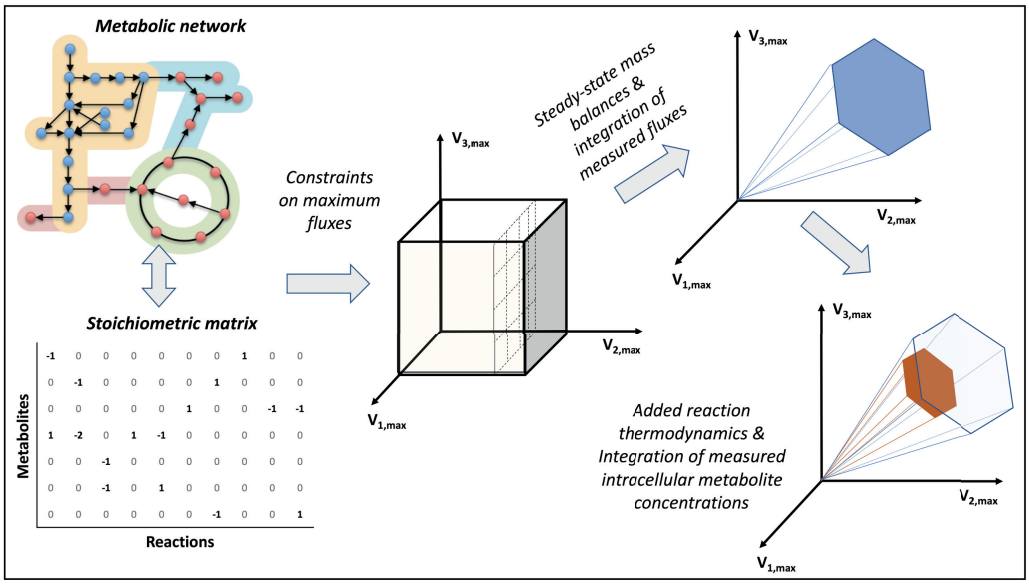


Figure 3.1 Constraint based modeling (CBM) formulation. A metabolic network is converted into a mathematical formulation that allows for the integration of constraints that limit the space of feasible flux solutions, thus constraining the metabolic phenotypes that are relevant for the physiological conditions under study.

When GEMs are reformulated into the aforementioned mathematical formalism they can become useful for the prediction of relevant metabolic pathways for certain physiological conditions. Their GPR annotation can be used for the study of gene to reaction essentiality studies, where the knockout of a single reaction is evaluated in the context of cell survival. Synthetic lethality studies can also be performed through the double knockout of reactions in the network (106, 107). GEMs can also be used for *in-silico* metabolic engineering purposes. The design of strains to perform some desired metabolic task, such as the production of a byproduct in high quantities, is often useful for industrial applications. Useful information and predictions for strain design can be obtained through studies on reaction knockout, knock in of reactions or pathways integrated from foreign organisms, and perturbations on enzyme expression (108-113). However, predictability from GEMs on strain design based uniquely on these approaches can be misleading and not fully representative of the behavior of the strain in vivo. Here, approaches that use kinetic modeling based on GEMs, where metabolic enzyme regulation can be studied, are more valuable as they provide a quantification of the

organism behavior to perturbations on kinetic parameters and enzyme concentrations (114).

3.1.2 Constraint-based Modeling (CBM)

Constraint-base modeling (CBM) can be applied to a GEM mathematical formulation with the purpose of constraining the feasible space of solutions of the fluxes in the network instead of focusing on a unique solution. This allows the account of multiple cellular processes that may happen under different physiological conditions pertaining to the feasible solution space as constrained by the topology of the network and by the assumption of steady state for all the mass balances in the network. Lower and upper bounds for the fluxes are applied as constraints to set the initial allowable space of solutions, which can then be further constrained to represent the physiological conditions of the cellular process under study (Figure 3.1) by integrating available flux measurements (uptake/secretion rates and/or intracellular fluxes). The feasible solution space can be further constrained by integrating reaction thermodynamic constraints and intracellular concentration measurements for the metabolites in the network.

The COnstraint-Based Reconstruction and Analysis (COBRA) toolbox is a MATLAB based package that has been developed with the purpose of disseminating methods for the reconstruction of GEMs and for their analysis, such as flux balance analysis, gene essentiality analysis, minimization of metabolic adjustment, Monte Carlo sampling, and gap filling, among others (115, 116). Flux balance analysis (FBA) is a method capable of simulating the metabolism in GEMs upon constraining the space with physiological data that uses linear programming to compute a solution for the fluxes that fulfills the steady state condition $\bar{\bar{S}} \cdot \bar{v} = 0$, where $\bar{\bar{S}}$ is the stoichiometric matrix and $\bar{v}$ is the vector of fluxes for each reaction in the network. FBA can produce multiple feasible solutions for the steady state condition. In order to select a solution (or set of solutions) that is relevant for the physiology being studied, it is common practice to add an objective function to the mathematical problem being solved that represents the proportion of which a set of metabolites are expected to be produced. Since the first GEMs, and hence, the first metabolic engineering analysis performed, were constructed

around organisms such as bacteria, parasites and some eukaryotes, the most used metabolic objective is the optimization of cellular growth. Growth requires the production of lipids, proteins, nucleotides and other essential compounds in specific amounts that fulfill the individual growth of an organism given the conditions determining their cellular uptake and secretion rates. This metabolic objective is formulated as a reaction, the *biomass reaction*, where the stoichiometric coefficients of each metabolite present in the reaction represent their proportions necessary to attain the growth requirement. Other metabolic objectives can also be formulated depending on the problem and the physiology under study, such as energy yield (ATP) and reductive power (NADPH/NADP+).

3.1.3 GEMs unification and annotation standardization

GEMs have been used for studying variability across different strains of an organism (117), host-pathogen interactions (118-120), gut microbiome interactions (121, 122), and in the study of mammalian/human metabolic diseases, such as non-alcoholic fatty liver disease (6, 123), diabetes (124) and cancer (7-9, 18, 125, 126).

The number of available genome scale metabolic reconstructions has seen a fast growth in the recent years for both prokaryotic and eukaryotic organisms, such as Saccharomyces cerevisiae, Pseudomonas stutzeri, Salmonella typhimurium, Staphylococcus aureus, Escherichia coli , Mycobacterium tuberculosis, Bacillus subtilis, Toxoplasma gondii, Plasmodium falciparum (118, 127-134), among many others. Following the same trend, genome scale models of human metabolism have been reconstructed and continuously updated on their reaction complexity:

- Recon 1(10) had its initial network derived from combining genome and enzyme data from different database sources, such as the Kyoto Encyclopedia of Genes and Genomes (KEGG) (135, 136) and EntrezGene (136, 137);
- The Edinburgh human metabolic network (EHMN) (138) was reconstructed from human enzyme gene information from different databases;

- HepatoNet1 (139) was reconstructed from aggregating the enzymatic reactions in KEGG and Recon 1, and manually curated to keep only a subnetwork for which there is biochemical evidence off its presence in hepatocytes;
- The Human Metabolic Reaction database (HMR) (12) was initially reconstructed for Adipocyte cell type by merging HepatoNet1 and pathway information from Reactome (140), which was then combined with Recon 1 and EHMN. Later it was expanded to HMR2 by incorporating further complex lipid metabolism (6);
- Recon 2 version 4 is currently the most updated version of Recon 1 (11), assembled by combining information from HepatoNet1 and EHMN and a literature based search on transport reactions.

Despite the improvements in the generation of high-quality reconstructions with the design of protocols to guide the procedure (141), the development of resources for building reconstructions, such as the ModelSEED (142) and The RAVEN toolbox (143), as well as the introduction of the Systems Biology Markup Language (SBML) format, which has contributed to make these models transferable (144), it remains difficult to compare reactions and compounds in GEMs (sometimes even of the same organism) when these are produced by different labs and through the use of different database resources during their reconstruction. Although recommended in the proposed workflow for reconstructions, not all reconstructed GEMs have their genes, reactions and compounds annotated and linked to external databases.

Poorly annotated GEMs bring limitations to their use, which require extensive work from the user of these different metabolic networks to match compounds and reactions. Firstly, failure to properly match metabolites and reactions across GEMs renders their comparison difficult, especially when the user is interested in understanding the subtlety of the network differences between GEMs of the same organism. Secondly, poorly annotated GEMs also limit the automatization of the process of integrating fluxomics and metabolomics data, and reaction thermodynamics parameters into the models, which are needed for constraint-based analysis. Inconsistencies or common errors found in GEMs, which are worsened by their lack of proper annotation, are:

- Erroneous naming of metabolites and reactions;
- Reaction and compound name format tailored for the particular use of the GEM at the time it was created. Examples are the use of in-house identifiers or the appending of chemical formulas and other identifiers to the names;
- Use of abbreviations that do not match accepted or listed compound and/or reaction names or identifiers

The problems mentioned above render automatized database searches difficult in order to remap the GEMs to database identifiers that assist with data and parameter integration. There has been an effort in creating model repositories that hold collections of GEMs with similar format and annotated through links with known databases. Such is the case of ModelSEED and BiGG Models (145), which at the time of the latest publication contained more than 75 annotated GEMs. However, these repository offers only a limited selection of consistently annotated GEMs and not a solution for remapping new problematic GEMs. Recently, BiGG has made an effort to re-annotate older models and provide them in SBML format.

There have been some attempts of unifying metabolite and reaction nomenclature with the purpose of facilitating GEM unification and providing assistance in reconstruction procedures. All these attempts use a similar approach for resolving metabolites and reactions. Briefly, they merge all compound and reaction information from selected databases through an iterative procedure where metabolites are matched primarily based on structural information (SMILES, InChI, etc...) and secondly by name and formula when structural information is not available. Reactions are resolved by matching participating substrates and products. Metabolites that have not been resolved are later on tested for matching in the context of reactions. BKM-react (146) resolved 20416 metabolites and 27367 reactions by merging three major databases (KEGG, BRENDA (147), MetaCyc (148) and now recently SABIO-RK (149)). MetRxn (150) currently accommodates 44783 and 35473 resolved metabolites and reactions, respectively, as a result from merging BiGG, BKM-react, BRENDA, ChEBI (151), HMDB (152), KEGG, MetaCyc and 90 GEMs. MetaNetX (153) was the latest attempt with the highest obtained number of resolved metabolites (82890) and 23210 resolved reactions

by merging the databases mentioned before and also LipidMaps (154), SEED, BioPath (155), Rhea (156) and UniPathway (157).

One limitation of these resources is that they can become outdated if not updated frequently by tracking the updates of the databases used in their construction. Currently BKM-react remains up-to-date with the last version dated from January 2017. However, at the time of this writing, MetaNetX namespace latest update was September 2015[1] and MetRxn last update was in 2014. Among other limitations are:

- model (and hence metabolite) mapping is biased to a search that is primarily focused in finding similar reactions in other models and databases and not in finding the most informative entries with respect to a compound identity, such as its structural information;
- since they are GEM unifying tools, typically a GEM format in SBML, COBRA for MATLAB, formatted excel, etc., is required to use these tools and not a list of names for instance;
- the user interfaces allow uniquely for the search of one compound or reaction at a time and not a list of names and do not provide web services for iterative search;
- the format of the output of the GEMs to be mapped can be inconvenient: in the case of MetaNetX it is returned a model structure that is uniquely mapped to their namespace identifiers losing all initial model information, and compounds and reactions for which a match has not been found are removed.

3.1.4 Implementation of dynamic pipeline

In this chapter, we focus on establishing a pipeline with the purpose of assisting with the mapping and identification of compounds in GEMs, for reconstruction and analysis, by using the web services directly provided by the compound and reaction databases. For this purpose, we developed the Data Repository for Automated Metabolomics Administration (DRAMA).

[1] At the time of the presentation of this thesis MetaNetX latest update occurred on July 2017, however, all results have been produced using the version of September 2015.

DRAMA framework was primarily developed with the purpose of facilitating data and parameter integration into GEMs. For instance, the analysis of metabolic models require that a lot of manual work is performed in order to match the metabolomics data to the corresponding metabolites. This work can become easily cumbersome when one is manipulating many different datasets from different sources, where metabolite nomenclatures seldom match between data sources and the GEM, which is worsened when identifiers from different databases or compound synonyms are also not provided. In addition, this pipeline complements the search for compound structure information that is essential for Thermodynamics-based Flux Analysis (TFA) and for our metabolic network reduction procedures, as will be seen in Chapter 4.

DRAMA aims to provide means for a flexible GEM annotation along with a data and parameter integration pipeline, which is independent of model curation. To ensure that we always recover the most updated information we use as much as possible the direct and online access to the selected databases through the web services provided by them when available. One further advantage of this pipeline is that all resources can be obtained without needing licensing or registration and they are independent of having a GEM structured as a model. The developed methods can also be integrated in a way that allows for a continuous and automatic update of internal lab databases. Furthermore, this pipeline is flexible enough that it can be exploited without needing a centralizing unique identifier. Parsing of search results and compound name preprocessing prior to search are the only things that can differ between implementations of the same approach and can be tailored to the specific needs of the user of these services.

The pipeline described in the chapter is applied to the human metabolic network Recon 2 v4 and sets the protocol for the standard operating procedure to assess and [re]-curate any GEM with the purpose of performing analysis that included reaction thermodynamics and for consistently generating condition-specific reduced metabolic networks that are tailored to the physiology under study.

3.2 Materials and Methods

3.2.1 DRAMA framework

The overall DRAMA pipeline (Figure 3.2) is developed in MATLAB programming language and consists of three main working stations: i) the GEM and dataset preprocessing, which uses GEMap pipeline that will be detailed in the next section, ii) the DRAMAcore repository, and iii) the AGADOR mediator. Both datasets and GEMs are preprocessed by calling adjacent pipelines that parse the metabolite names and collect any available accompanying external database identifiers for compound identification and matching. After the preprocessing the GEMs are stored for later use and the datasets are added to the repository of data in the DRAMAcore.

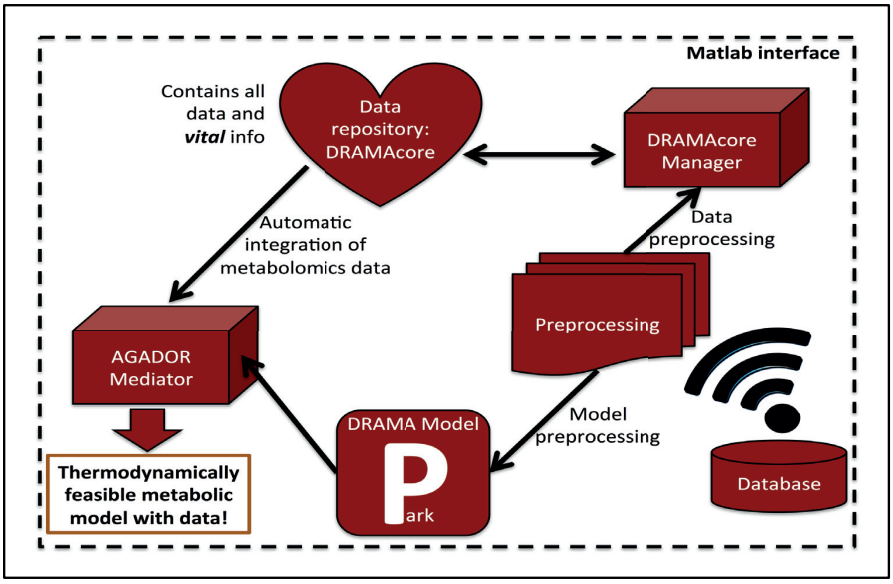


Figure 3.2 Schema of DRAMA main working pipelines for GEM annotation, data integration and analysis.

The DRAMAcore is a dynamic data repository that gathers fluxomics and metabolomics data from different organisms in a compact and organized way (Figure 3.3). Data is organized by cell types, tissue, conditions, disease state, and others. Its detailed annotation also allows for tracking data sources and experimental conditions improving citation retrieval. Datasets are curated by the users into a mandatory predefined format in excel that can then be uploaded to the repository. Each fluxomics and metabolomics dataset contains metabolite names and external database identifiers (if provided by the experimental sources) that are used for compound matching during

or prior to upload. Internal (optional) and external identifiers are associated with each compound, which will be used for matching the data to the GEMs.

Since analysis with GEMs using COBRA toolbox is typically performed in flux units of mmol/h/gDW, each fluxomics dataset is stored in this repository in that unit or mmol/h/cell. The user responsible for uploading the dataset should provide a field for conversion into mmol/h/gDW containing the cell weight and the respective reference. The same is valid for the metabolomics datasets where the concentrations are provided in mol/L or mol/cell with respective cell volume for conversion. The conversion factors provided are used as default values. Approximations for similar cell lines can be used if real measurements are unknown and this should be indicated as such in the appropriated information field. The pipeline in AGADORmediator (see below) allows the user to choose to use the default conversion factors or provide new ones.

The purpose of this structure is to facilitate data selection and sharing across collaborators, minimizing manipulation by users, which could introduce errors. Although it is presented here (Figure 3.3) as a MATLAB structure, because it is the main software used in our analysis, it can easily be converted into a database format for use across platforms, which is ultimately the goal. Users can also choose to create independent data repositories, using the pipeline to build DRAMAcores specific for the project they are working on for data that cannot be shared.

The AGADORmediator station contains functions that guide the user through the data available in the data repository (selected DRAMAcore) and preparation of the GEM for analysis. It is the final stage of the DRAMA main pipeline. Upon selection of a GEM and the datasets corresponding to the phenotype/condition to be studied, the metabolomics and fluxomics data are converted into the final units and into lower and upper bounds that can applied to the GEM. The overall procedure is summarized in Figure 3.4. The user can select the datasets to be used making use of the system of tags in the repository (organism, tissue, phenotypes, condition, etc, ...) or can select individual datasets by their positioning in the repository or by article reference. Specific datasets can be excluded by selecting properties not to be applied; for instance, the user can select all

conditions except hypoxic types or exclude a dataset from being selected by its reference.

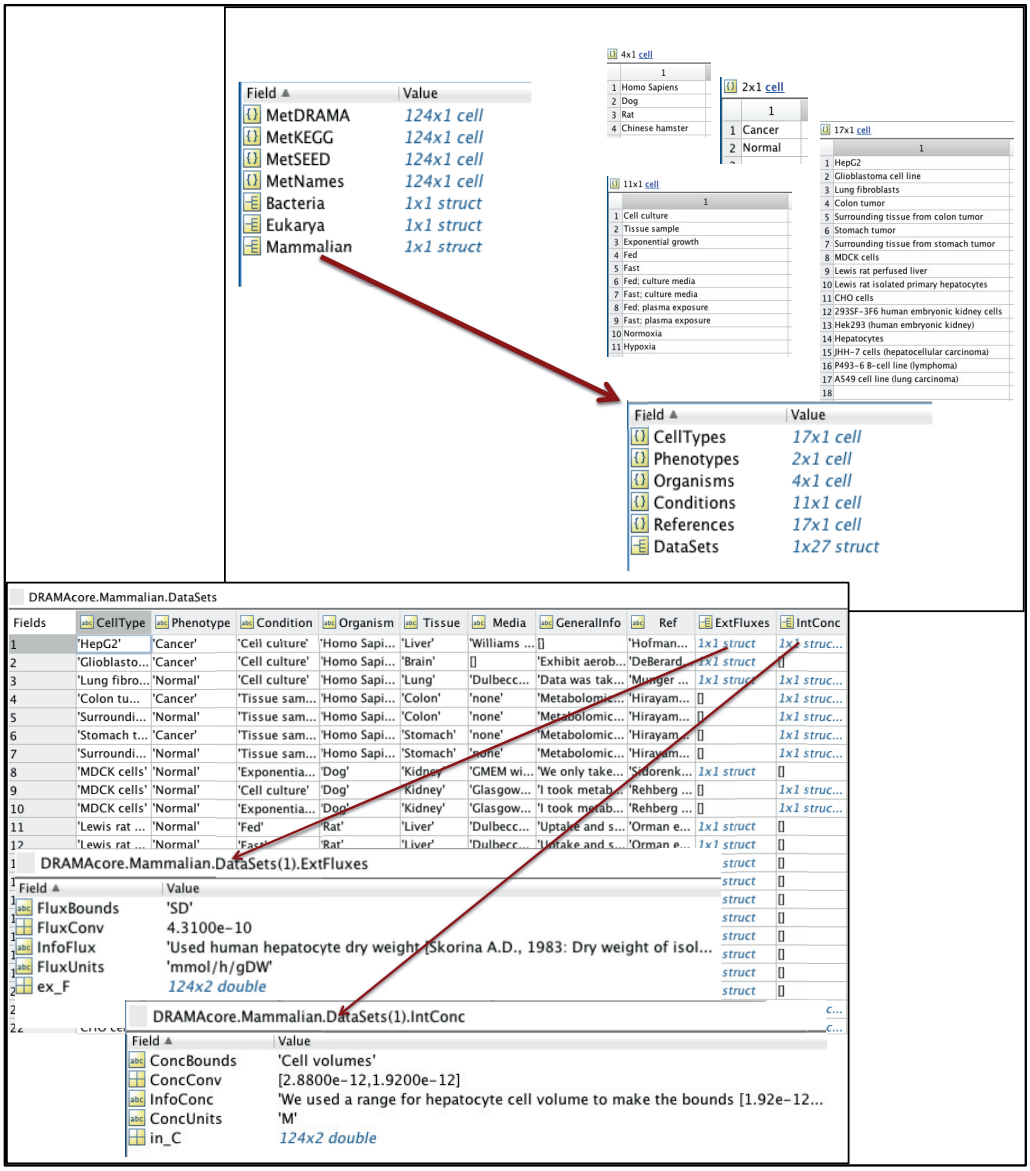


Figure 3.3 Visualization of format and content of DRAMAcore repository, here presented as a MATLAB structure.

After this step, the user can further decide about conversion factors to be used for cell volumes or cell weights for metabolomics and fluxomics data, respectively, as

well as providing instructions for the computation of the lower and upper bounds to be applied. Lower and upper bounds are computed through user specifications regarding how much to be subtracted and added, respectively, to the measurements in the dataset. These instructions could be to use the standard deviation (SD) of the measurements in the original datasets, use SD multiplied by a factor for further relaxation, relaxation by a percentage of the flux or intracellular concentration value (10%, 20%, ...), etc.

The metabolites in the repository are matched through their identifiers to the identifiers of the metabolites in the GEM (assuming it has been properly preprocessed with GEMap as seen below). Intracellular concentration constraints are applied directly to the bounds of the respective intracellular metabolite at each cellular compartment and data constraints for cell uptakes and secretions are applied to the GEM if it contains extracellular reactions involving the metabolites present in the extracellular fluxomics dataset.

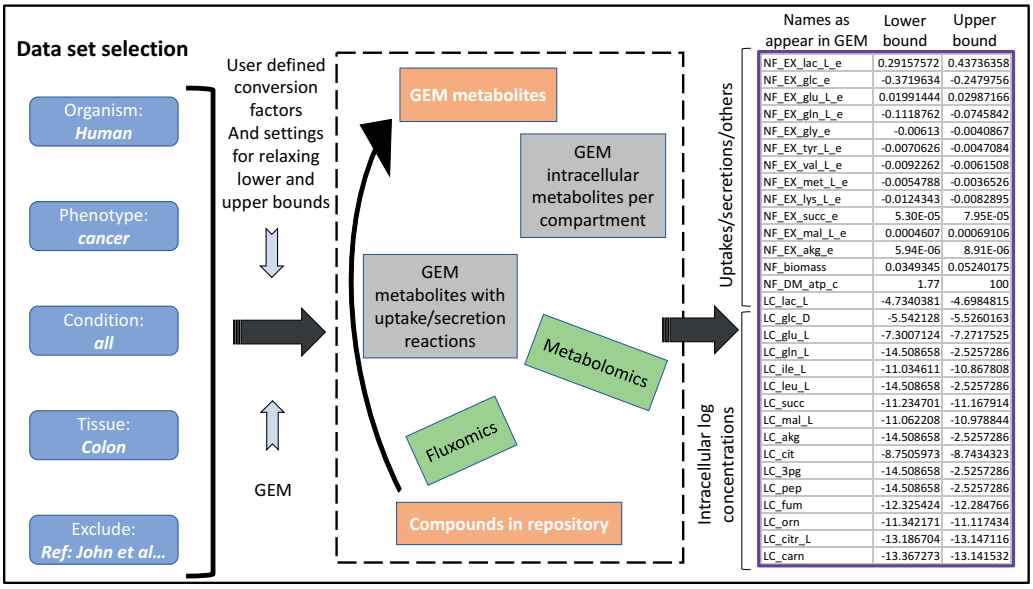


Figure 3.4 Summarized procedure for data integration into the model. Datasets are selected through identifier tags and ranges for flux and intracellular concentration bounds for constraint-based problem are applied to the model. Extracellular reactions are identified by matching the identifier of the participating metabolite in the GEM to the one in the dataset.

3.2.2 GEM & data preprocessing

The preprocessing station harbors GEMap, a pipeline that works within MATLAB connecting with multiple databases simultaneously. Through this pipeline, we favor connections to the databases via their provided web services whenever available. Web services enable access to a variety of data through HTTP requests that are embedded in third-party applications (such as MATLAB or R) or other scripts, which enable programmatic access that can be carried out in an iterative way. Table 3.1 contains a listing of the databases accessed within GEMap and the protocols used for programmatic access (see section A.3.1 for more description of the web services protocols and systems requirements for implementation). Currently the pipeline connects seven databases, two of which (SEED and HMDB) are not accessed online as they do not provide web services (see section A.3.2. for their set up as local databases instead). Programmatic access through the use of these web services allows us to have the most updated information in real time and avoids the exhausting process of constantly updating our local generated SQL databases with the newest releases.

With GEMap we can: (i) obtain external database identifiers given a list of "sensible" compound names if no other identifiers available, (ii) retrieve the most updated compound structural information by searching external databases through the identifiers associated with the compounds based on compound name search or database identifier search, (iii) propagate the list of external identifiers by accumulating cross-reference identifiers among databases. The input consists on a list of compound names and/or external database identifiers, which could both be obtained from a GEM or a metabolomics dataset. The fact that the input can be a list allows for flexibility in the uses of this pipeline, where a complete GEM structure is not necessary (we do not evaluate reaction information). GEMap can operate in two modes, which have different purposes: 1) all databases are simultaneously queried based on the provided list of compound names and/or database identifiers (Figure 3.5) or 2) search is oriented to matching compound names and external database identifiers to the identifiers of a particular database (Figure 3.6).

Mode 1 is used to match the compound names and external identifiers queried to as many cross references as possible. The queries for each compound in the list run in parallel, but within the same compound search, the databases are accessed in series.

The process can exectute faster by giving priority to the cross referencing. For instance, if a match is found in ChEBI database for the compound being searched, one can immediately reuse all the existing cross references to the other databases included in our search and skip the search in those. This avoids sending numerous requests and speeds up the process.

Mode 2 is used to match compounds to a particular database and further decreases the number of unnecessary requests sent. The search is conducted first using the connection to the database of interest (for instance, KEGG as shown in Figure 3.6). In a second step, the compounds that have not been matched in the primary search are collected and sent as requests to the remaining databases. If a match is found in one of these databases that bears a cross reference to the database of interest a "reverse mapping" is applied where the KEGG id is selected for the mapping. As shown in the example, C00031 is selected for the mapping based on the search CHEBI:4167. Once a reverse mapping is found the search is finished for that compound. If after "reverse mapping" there are still compounds the have not been matched to the primary database (KEGG in this example), they will have all the other database ids found during the search collected in order to ensure some identification for that compound ("assisted mapping" results). See LMFA01050442 case for example.

Table 3.1 List of databases accessed within GEMap.

Databases	**Access**			
	REST[2] API[1]	*SOAP[3] API*	*DB at local server*	*URL*
Kyoto Encyclopedia of Genes and Genomes (KEGG)	X			http://www.kegg.jp
PubChem	X			https://pubchem.ncbi.nlm.nih.gov
Lipidomics Gateway (LipidMaps)	X			http://www.lipidmaps.org
BiGG[*]	X			http://bigg.ucsd.edu
Chemical Entities of Biological Interest (ChEBI)		X		https://www.ebi.ac.uk/chebi/
MetaCyc	X			https://metacyc.org
ModelSEED (SEED)			X	http://modelseed.org
The Human Metabolome database (HMDB)			X	http://www.hmdb.ca

[1]API (Application Programming Interface) consists of a set of procedures that allow the access to data.
[2]REST: Representational State Transfer
[3]SOAP: Simple Object Access Protocol
[*]BiGG database server is accessed through a localhost for speed

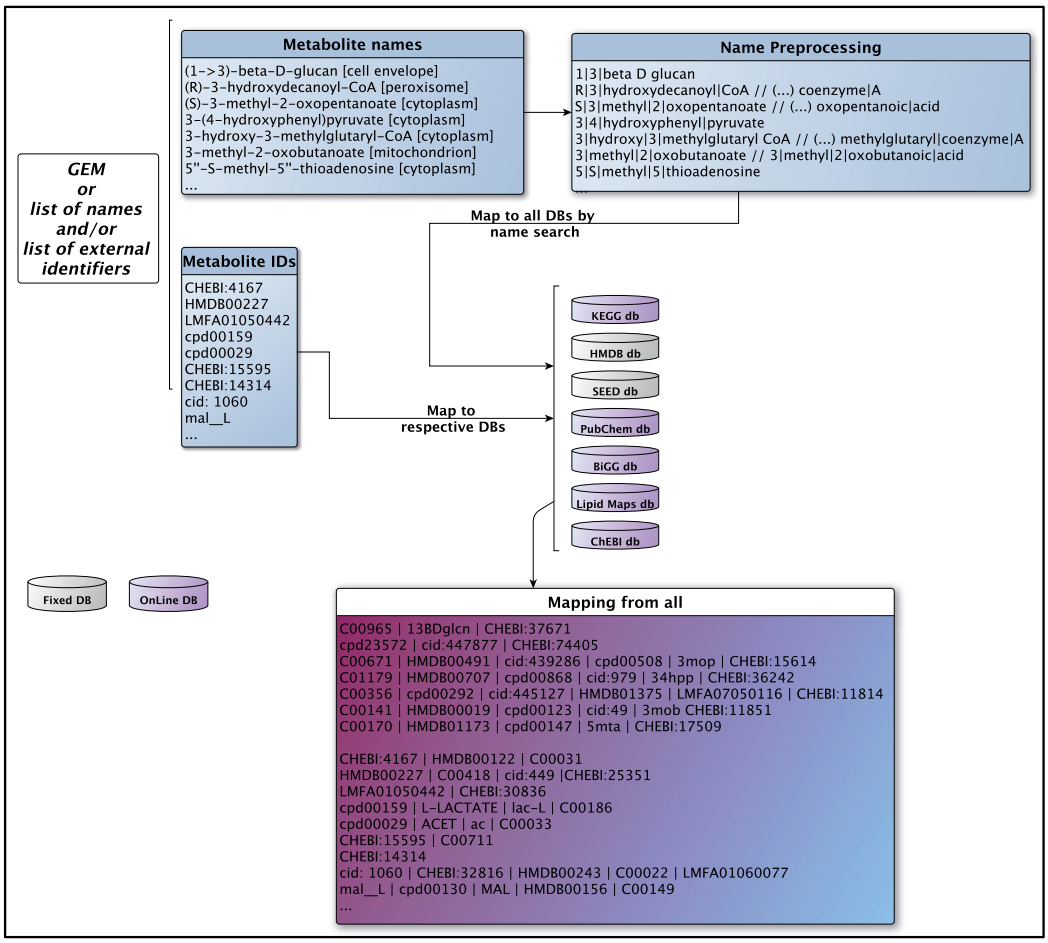


Figure 3.5 GEMap pipeline is shown here with the search in mode 1 where the purpose is to match the compounds to as much external database identifiers as possible. As input a GEM or list of names is provided alongside a facultative list of existing external database identifiers. Databases highlighted in grey have no web services and are installed in local servers. Note: Metacyc is not included in the diagram because its web services do not allow search by compound name, just internal and external compound identifiers (see Table C.3.1 for a summary of web services provided per database).

Compound names are queried by removing only the prime symbols (5',3', etc...). In its most restrictive, a match is retrieved if a full string match is found between the search name and the synonym names provided in the database. For string matching, further processing of the compound names is used as shown in the GEMap diagrams where all non-alphanumeric characters are removed (except for + and -sign associated with charges) and subsequently replaced with vertical separators '|'. This name processing is performed post search in both the searched compound name and the names from the search results, which include compound generic name and their

associated synonyms. Since chemical nomenclature is known for its diversity, as shown by the focus of several studies on strategies to improving string matching algorithms (158-160), this quite simplistic processing step ensured a fairer string matching comparison of all groups in the compound name independent of their order of appearance. During mapping in mode 1, if synonym names are not provided in one particular database, such as in the case of BiGG, or they are limiting, InChI (removing protonation state) and external database identifiers can be cross-checked with the results from the previous database requests for that compound, which run in series.

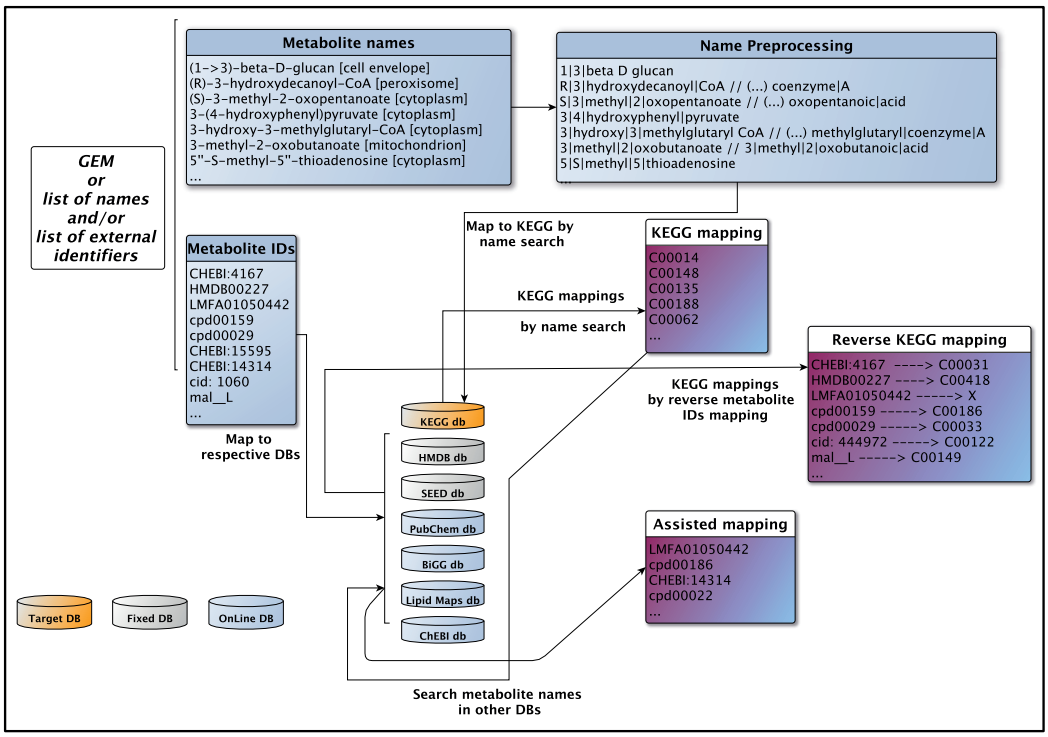


Figure 3.6 GEMap pipeline is shown here with the search in mode 2 oriented to KEGG compound identifiers (database selected in yellow). As input a GEM or list of names is provided alongside a facultative list of existing external database identifiers (other than KEGG). Search is done accessing KEGG web services first. In a second stage assisted mapping starts by looking up the compounds not matched within the other databases for reverse mapping to KEGG identifiers using cross referencing. Databases highlighted in grey have no web services and are installed in local servers. Note: Metacyc is not included in the diagram because its web services do not allow search by compound name, just internal and external compound identifiers (see Table C.3.1 for a summary of web services provided per database).

3.2.3 Compound thermodynamics curation

A molfile contains the information regarding the structure of a molecule describing its atoms, their bonds, connectivity and positional coordinates, which can be downloaded from the compound databases. For GEMs containing networks linking unique metabolites in the order of thousands it can rapidly become a cumbersome task to manually search for these compounds in the different databases, download the respective molfiles and save the files with formatted names that facilitate iterative import into MATLAB (or other programming tools) for further computations. Furthermore, since not all databases will have the desired compound and some may even contain unreadable molfiles, it is imperative to search different databases in parallel.

At this stage, the compound mapping results, obtained either from GEMap operating at mode 1 (Figure 3.5) or from other resources, which link metabolites in the GEM to identifiers of multiple databases, can be readily used as input back to the databases listed in Table 3.1. This is done through requests that specifically retrieve the associated molfiles in a *.txt* format that can be easily stored (Figure 3.7). Simultaneously, the molfiles readability is tested by converting them into InChI structures using the molconvert functionality in, Marvin 16.7.4, 2016, ChemAxon (http://www.chemaxon.com) (161). It is at this stage that the collection of multiple external database identifiers performed earlier plays a role. Since not all databases contain structural information for a compound, this propagation of identifiers ensures a higher number of molecular structure retrievals. Results from the search can be stored as individual molfiles in a folder with appropriated formatted name for import at later stages, or can be preserved into the GEM model structure (in MATLAB readable format, for instance). Note that BiGG database does not provide content for compound structural information (see Table C.3.1 for a list of web services provided per database).

This pipeline is integrated with the existing pipeline in the lab where molfiles are used for estimating the standard Gibbs free energy change of formation ($\Delta_f G^\circ$) for a metabolite, which is the change in energy required to form 1 mole of the substance under standard conditions of pressure and temperature. This quantity is used to estimate the standard Gibbs free energy change of reaction ($\Delta_r G^\circ$), which can be used to assess the reversibility of reactions in a network and further constraint the space of

feasible flux solutions in CBM. $\Delta_f G^\circ$ is estimated using the Group Contribution Method (GCM) published previously (162). Briefly, GCM estimates $\Delta_f G^\circ$ by decomposing the molecular structure of the compound in the molfiles into smaller substructures (groups) and adding the energy contributions of the different groups forming the compound. For the computation of $\Delta_f G'^\circ$ in aqueous solution, the compound predominant ionic form used is the one at pH 7. The charge of the compound at pH 7 is obtained using the pKa values estimated with MarvinBeans 16.7.4, 2016, pKa calculator in cxcalc, ChemAxon (http://www.chemaxon.com). The $\Delta_r G'^\circ$ can then be estimated by subtracting the sum of the $\Delta_f G'^\circ$ of the substrates/reactants to one of the products in the reaction

$$\Delta_r G'^\circ = \sum \Delta_f G'^\circ_{products} - \sum \Delta_f G'^\circ_{substrates}. \quad (3.1)$$

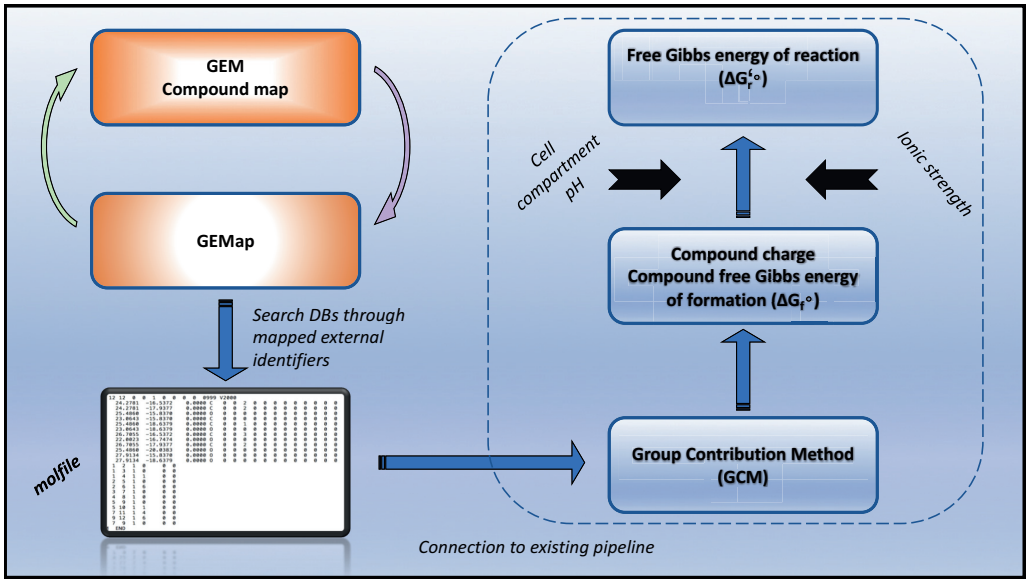


Figure 3.7 Schema of DRAMA pipeline integrating GEMap results (or any other available compound mapping) and the service to request and automatically download molfiles by searching through the compound external identifier in the respective databases. The pipeline follows with the subsequent estimation of the $\Delta_r G'$.

The Gibbs free energy change of reaction ($\Delta_r G'$) depends on the concentration of the metabolites involved in the reaction. However, each metabolite has a solution-dependent activity, which is affected by the ionic strength of the reaction medium, that

reflects the real amount of active compound available for the reaction, and also by the pH inside the cellular compartments that affects its charge. We take all these cellular conditions into account (see section 3.2.6 for the values used) when we estimate the corrected $\Delta_r G_j'^{\circ}$ of each reaction j in a GEM. We compute the metabolite activity in function of the ionic strength using the extended Debye-Huckel equation (163), which is then absorbed into the estimation of $\Delta_r G'^{\circ}$. The $\Delta_r G_j'$ of a reaction j is then estimated by using the equilibrium constant, which is a function of the concentrations of substrates and products in the reaction, as follows

$$\Delta_r G_j' = \Delta_r G_j'^{\circ} + RT \cdot \sum_{i=1}^{n} n_{i,j}\, ln(c_i), \tag{3.2}$$

where R is the ideal gas constant (8.31446 J mol-1 K-1), T is the temperature at standard conditions (298 K), $n_{i,j}$ is the coefficient of the stoichiometric matrix for metabolite i participating at reaction j, and c_i is the concentration of metabolite i in the cell.

3.2.4 Thermodynamics-based Flux Balance Analysis (TFA)

In section 3.1.2, we introduced the concept of Constraint-Based Modeling (CBM) and Flux Balance Analysis (FBA). The space of feasible solutions in the analysis of a GEM with fluxomics data integrated pertaining to a specific condition can be further constrained by adding reaction thermodynamics and metabolomics data, such as intracellular metabolite concentrations, as shown in Figure 3.1.

Reaction thermodynamics helps to determine the directionality of the reactions in the GEM. A $\Delta_r G'$<0, according to the second law of thermodynamics, indicates that the reaction is carried in the forward direction. Thermodynamics-based flux balance analysis (TFA) is a re-formulation of the FBA problem presented in section 3.1.2, and was initially presented in (164). In the initial FBA formulation, the reactions in the network are mathematically represented by the stoichiometric matrix S and are mass and energy balanced. The physiological condition under study is simulated by computing a solution that fulfills the steady state condition $\bar{\bar{S}} \cdot \bar{v} = 0$, where $\bar{v}$ is the vector of fluxes for each reaction in the network and is constrained by imposing upper bounds in v_{max}. TFA adds a set of mixed integer linear constraints to this problem along with eq. 3.2 for the reaction thermodynamics, allowing for the problem to be dependent

on the concentration level of metabolites reported/measured for the physiology and conditions being studied.

TFA formulation allows for a flux-balance analysis of the system in a way that removes infeasible reactions or pathways from flux profile solutions, and at the same time provides an insight to the ranges of metabolite concentrations that determine the directionality and feasibility of a reaction. We can then define objective functions and use optimization algorithms to determine the flux profile distributions that lead to the maximum growth yield, the highest net production of a desired byproduct, or the cellular responses to gene deletions (reaction knock outs), as well as responses to incorporation or pathways/reactions foreign to the organism.

3.2.5 GEMs used in this chapter

The GEMs used are iMM1415 (165), the *Mus Musculus* mouse model and Recon 2 v4 (11), the generic human metabolic model. The version of iMM1415 used was the one obtained at a time closer to the model publication where the compound KEGG annotations were obtained upon request to the authors.

3.2.6 Physiological mammalian/human cell parameters

For the estimation of $\Delta_r G'^{\circ}$ we use the cell physiological parameters for pH and ionic strength in Table 3.2. These parameters where obtained from literature search on mammalian cell physiology. The value of cytosolic ionic strength used in our estimations was 0.15 Mm, which was reported for mammalian cells (166), and within the range of values to apply the extended Debye-Huckel equation (167) (up to 0.35 M). The pH values for each mammalian cellular compartment where compiled from multiple works reviewed in (168). We have also compared the $\Delta_r G'^{\circ}$ values estimated with ionic strength 0.15 M to the ones estimated with 0.25 M ionic strength, which was the value used for the estimation of reaction thermodynamics for about 2/3 of reactions in Recon 1 (169). Our results show that mostly very small deviations in the estimated $\Delta_r G'^{\circ}$ occur between these two ionic strength values (Figure B.3.1) and only five reactions with thermodynamic constraints, which are not transports, had a switch on the sign of $\Delta_r G'^{\circ}$ (Table C.3.2). However, taking into account the associated $\Delta_r G'^{\circ}$

uncertainty and the range of physiological concentrations, these five reactions are maintained as reversible in both cases.

Table 3.2 Mammalian cell physiological parameters. The pH values were obtained from (168). The ionic strength for mammalian cell cytosol was obtained from (166), cross checked with computation using the cytosol ion concentrations in (170, 171) and assumed equal for all compartments.

Cell compartment	*pH*	*ionic strength (M)*
Cytosol	7.2	
Mitochondria	8	
Vacuole	7	
Perixome	7	
Extra-organism	7	0.15
Golgi Apparatus	6.35§	
Endoplasmic Reticulum	7.2	
Nucleus	7.2	
Lysosome	4.7	

§ Average of pH across Golgi

3.3 Results and Discussion

3.3.1 Mapping compounds of GEMs

We mapped Recon 2 v4 with GEMap using the unique compound names provided in the model. Table 3.3 summarizes the statistics on the total number of compounds for which a match is retrieved during search by compound name in the databases. As explained in section 3.2.2, a match is accepted if all the substrings of the processed compound name used in the search fully matches the substrings of the processed synonyms from the database request output. The maximum number of matches retrieved per compound across the databases was 48.2%, which was obtained for HMDB. HMDB does not have web services for online search and it is stored locally. Right below with 46.2% is PubChem, which is queried through the use of available web services. PubChem contains an extensive list of synonyms per compound, which improves our matching. The database with the lowest performance in matching compound names was Lipid Maps, with 8.5% of compounds matched. Lipid Maps is a database that focuses on lipids, which characteristically possesses much more complex nomenclature that is thus

harder to match with block full string matches, as the one applied here. The second lowest database matching performance occurred for BiGG, with approximately 21% unique compounds matched. This low match performance occurs because BiGG provides no synonyms for compound names and their format is often non-standard. However, included in this number of matches are also the records that even though didn't match in compound name, had external identifiers that intersected with previous records retrieved for those compounds in the other databases searches prior to BiGG.

We compared the compound name mapping with GEMap to two other sources to assess information complementarity: MetaNetX[2] (MNX) (153, 172) and the compound identifiers existing in the GEM structure (KEGG, ChEBI, PubChem, HMDB). Figure 3.8a presents the statistics on compound mapping coverage compared to GEMap results. The compounds in the GEM are classified by having been attributed a database identifier by both GEMap and the other source (*Both GEMap and the other source have result*), by having been mapped by GEMap but not by the other source (*GEMap result but not other source*), by having been mapped by the other source but not by GEMap (*Other source has result but not GEMap*), and by having not been mapped (*Neither has result*). In the overall, GEMap search by compound name was able to identify 70% of the unique compounds in Recon 2 v4 network, whereas MNX identified about 80% and Recon 2 v4 IDs covered about 63% (Figure 3.8b). Both GEMap and MNX were able to provide more coverage than the one pre-existing in the GEM, with MNX achieving the highest level. However, GEMap shows an advantage on complementing compound mapping and identification, which can be useful to provide more structure information for the estimation of reaction thermodynamics in the models. In addition, GEMap has been observed to complement compound identification with database links that contain molecular structures when MNX mapping provided only BiGG identifiers that have no compound structure information (a subset with an example can be found in Table C.3.3

[2] The model was uploaded in SBML format into MNX web interface (http://www.metanetx.org/cgi-bin/mnxweb/mnet_upload) and the results file in excel format was parsed for analysis. MNX namespace combines information from multiple compound and reaction databases and maps the GEM by identifying the metabolites in the context of the reactions present in the network. Since its main purpose is to unify GEMs, the resulting network output in SBML is formatted with their MNXidentifiers and contain only metabolites and reactions that have been properly matched and identified to their namespace. Exporting the results in the excel file format, which contains information about the original model, the non-matches and the identifiers of external databases is more useful than the SBML one, however, it requires more parsing steps.

green and blue section). This can be a consequence of MNX algorithms whose purpose is model unification by reaction identification, thus settling for the information on the database for which the reaction matched, in this case, BiGG.

We analyzed how many mapped compounds did not have common external database identifiers between GEMap-MNX (160 compounds) and GEMap-Recon 2 v4 IDs (140 compounds). This was done by intersecting the external identifiers from the individual database searches obtained for each compound in GEMap with the external identifiers resulting from the mapping in MNX or the pre-existing Recon 2 v4 IDs. Note that this number does not directly indicate inconsistency or error. The fact that no intersection is found can be primarily linked to: lack of cross-referencing, selection of stereoisomers or tautomers as a consequence of the search based on compound names.

The 160 compounds that did not have common external database identifiers between GEMap-MNX are listed in Table C.3.3. A deeper inspection revealed that the lack of common identifiers is mostly due to MNX linking unique compound identifiers to BiGG identifiers that have poor compound name identification and no external database cross-referencing. Indeed, for 85 of the 160 compounds GEMap was able to map these compounds to BiGG identifiers with the same name in MNX mapping, however, in our case they are CHO model associated (ex: pa_cho), whereas for MNX they are associated to the homo sapiens counterpart (pa_hs). In this case, even though the compound was the same, the result was scored as a mismatch because 33 of the 85 compounds had no other associated external identifiers for both GEMap and MNX. For the remaining compounds of the 85 set, GEMap provided external database identifiers other than BiGG. Since in its base MNX merges reaction and compound information of several GEMs, we speculate that those BiGG identifiers remain separate as they may have been observed to occur with those names only for a specific reaction pertaining to a BiGG model.

Another set of 44 compounds were also uniquely mapped to BiGG identifiers in MNX, whereas GEMap retrieved cross-referencing for other databases, except BiGG. For this set, the compound name in the model seems to be the cause of the mismatch and stereochemistry or protonation differences seem to be playing a role. Often, in databases, compounds are identified separately based on differences in stereochemistry

and protonation, whereas in other databases they can be found merged through synonyms. Since the purpose of our pipeline is mainly to propagate the available number of cross-references in order to ensure retrieval of compound structural information, these differences in protonation and stereochemistry won't influence our results as they have little effect on the estimation of $\Delta_f G^\circ$ and hence are decomposed in the same way during computations with CGM. Protonation is dealt with as the molecule structure is converted to pH 7 prior to estimations.

Finally, 23 compounds with inconsistent cross-references between the two mappings also presented the differences in stereochemistry or protonation referred above. Within these, 13 were uniquely mapped to BiGG with GEMap, which does not provide compound structural information. In the overall, only 6 compounds seem to present a more severe reason for mismatch of their cross-references and nomenclature inconsistencies may play a role. In fact, one of the 6 compounds, "20-hydroxycholesterol", appears twice in the list of unique compounds in Recon 2 v4 associated with different unique metabolites and different reactions. Other such case is dolichyl phosphate(2-), which is also associated with two separate BiGG model compounds (dolp__L and dolp_U) In such cases, contextualization of compound identification based on reactions becomes more useful.

Table 3.3 Statistics on the mapping of Recon 2 v4 compounds using GEMap at mode 1 by searching uniquely with compound names.

After individual DB search (without cross referencing)	**% matches per compound name search**
chebi	34.3
hmdb	48.2
kegg	29.4
lipidmaps	8.5
seed	36.7
pubchem	46.2
bigg	20.8

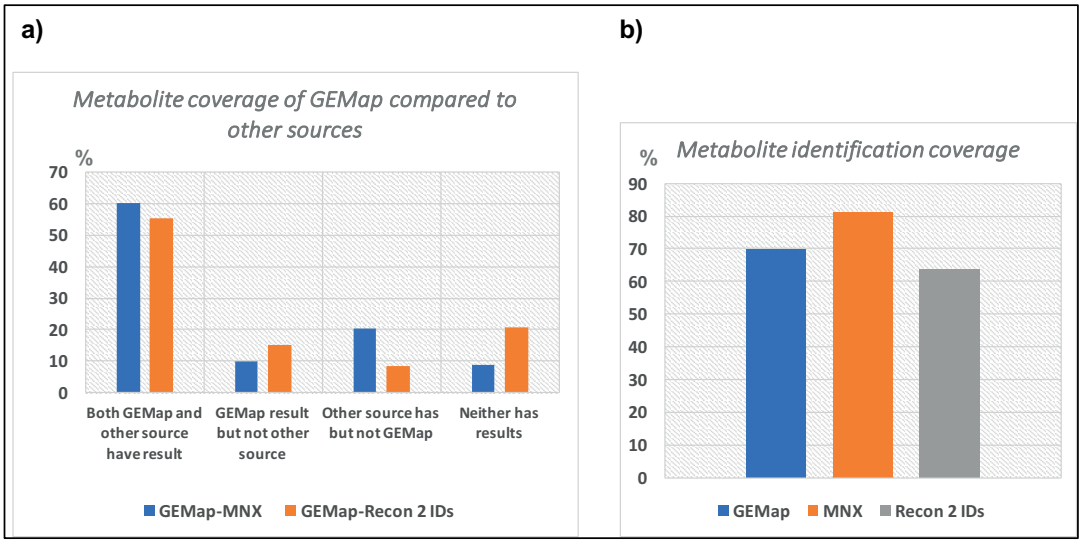


Figure 3.8 a) Comparison of metabolite mapping coverage in Recon 2 v4 between GEMap using compound name search and MNX results and between GEMap results and the existing compound identifiers in Recon 2 v4 structure. Statistics computed with respect to the number of metabolites that are or are not attributed an identifier with any of the three sources. b) Total metabolite mapping coverage using GEMap with compound name search, MNX and Recon 2 v4 identifiers.

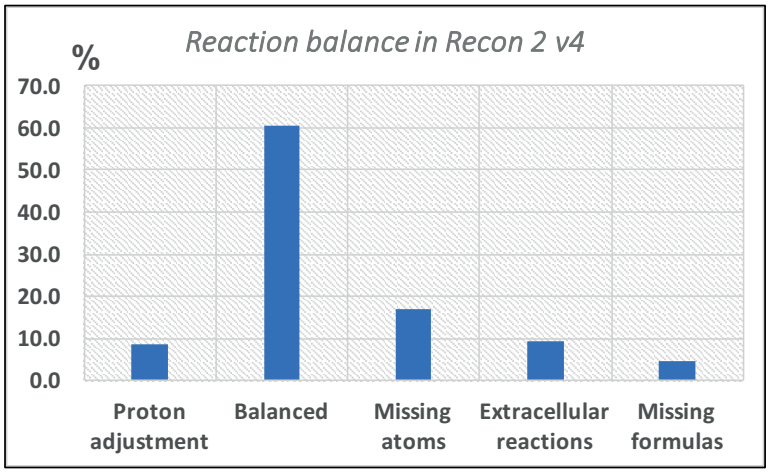


Figure 3.9 Statistics on number of reactions in Recon 2 v4 that were balanced, that required proton adjustment and that had missing structures and hence balance was not assessed.

3.3.2 Reaction balance and assessment of pre-assigned directionalities

After compound identification in Recon 2 v4 and extraction of the molfiles with the compound structure, the correct chemical formulas and protonation states for the

mammalian cellular compartment conditions (pH and ionic strength) can be derived. These are used to assess the balance of all chemical reactions in the model, except for extracellular reactions. This is a necessary step that needs to be taken prior to estimation of reaction thermodynamics.

Our model assessment showed that about 60% of the reactions in Recon 2 v4 were balanced. With the available compound structures, we were able to perform proton adjustment in 10% of non-balanced reactions. Most of the remaining reactions had structures with a unknown number of group repeats or R groups which could not be used to assess reaction balance and a small percentage had no chemical formula associated.

Recon 2 v4 as presented in (11) was curated starting from the previous human network reconstruction, Recon 1 (10), with 2/3 of reaction directionalities assigned resulting from a reaction thermodynamics study (169) and assembling network information from HepatoNet1 and EHMN as well as a literature based search on transport reaction. In order to prioritize our reaction thermodynamics curation and avoid blocking of reactions whose pre-assigned directionalities in the model are inconsistent with reaction thermodynamics, we investigate which pre-assigned reaction directionalities in the model differed from our thermodynamics curation (Figure 3.10). Reaction directionalities were estimated as described in section 3.2.3. Most directionalities that differed were pre-assigned in Recon2 v4 as forward (F) reactions, whereas our thermodynamics curation supported reversibility. We note, however, that this high proportion of thermodynamically reversible reactions results from the wide range of intracellular metabolite concentrations used in the calculations (set default as 1e-11 to 0.08M since human metabolomics typically contain measurements at 1e-11 M order of magnitude). Integration of metabolomics data in the model for a particular phenotype will help determine the directionality of these reactions. A much smaller proportion of the directionalities changed corresponded to reactions that were pre-assigned as bidirectional (BI) in Recon 2 v4 while being thermodynamically feasible only on the F or reverse (R) direction. Reactions blocked in Recon2 v4 where unblocked and set to their thermodynamically consistent directionalities. The five reactions that

were set as F but were R as per thermodynamics curation were searched in the literature for directionality information: three of the 5 were reversible in Brenda and the other two were only found in BiGG database with no other external database link to infer directionality.

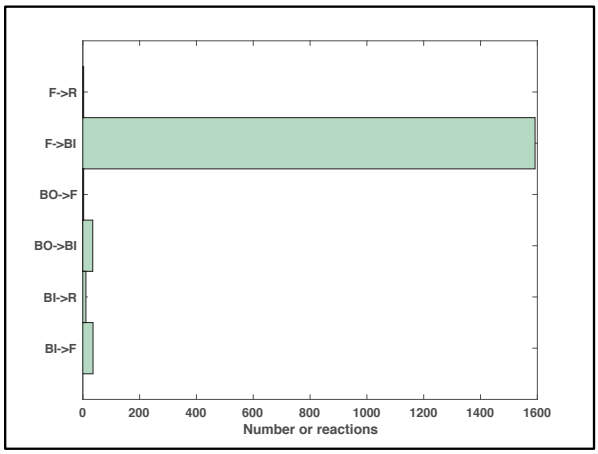


Figure 3.10 Number of reactions in Recon 2 v4 whose original flux constraint bounds were modified based on the directionality of the reaction within thermodynamically feasible ranges. BI: Bidirectional reaction (or reversible in case thermodynamics constraints are applied); F: Forward reaction; R: Reverse reaction; BO: Blocked reactions (carry zero flux).

After the thermodynamics curation of the pre-assigned reaction directionalities (Table 3.4) the proportion of bidirectional reactions was (65%) (column 1), which was higher than prior to the curation of the directionalities (Table C.3.4). Flux variability analysis (FVA) of the feasible flux space imposing only the mass-balance constraints of the network shows a reduction of about 24% bidirectional reactions (column 2), which is the same drop observed before our curation (Table C.3.4). Additional constraining by imposing reaction thermodynamics on this network without condition specific metabolomics data didn't decrease significantly the flexibility of the network, as shown by the thermodynamics-based flux variability analysis performed (TFVA) (column 3). Further constraining of the network relies thus strongly on the integration of condition specific metabolomics data for the analysis of the metabolic phenotype of interest.

Table 3.4 Statistics on reaction directionality at different stages of flux constraining for Recon 2 v4 after imposing directionality's based on reaction thermodynamics assessment for a wide range of physiological metabolite concentrations. Column 4 is just a statistic on the reactions from the model that have thermodynamics constraints.

Reaction directionality	*Directionalities imposed as flux constraints*	*Directionalities after evaluating the feasible solution space using only network mass-balances*	*Directionalities after evaluating the feasible solution space after applying thermodynamics constraints*	*Directionalities of reactions with thermodynamics constraints for default metabolite concentration ranges*
BI	4854 (65%)	3077 (41.3%)	2938 (39%)	3337 (91.1%)
F	2062 (27.7%)	2081 (27.9%)	2054 (27.6%)	313 (8.5%)
R	58 (0.78%)	256 (3.44%)	255 (3.4%)	14 (0.4%)
BO	458 (6.1%)	2028 (27.2%)	2195 (29.5%)	-

3.3.3 Thermodynamics curation gain by automatizing

Here we illustrate the thermodynamics constraints coverage for Recon 2 v4 and compare it to the one from the iMM1415 *Mus musculus* model. Both GEMs were mapped using MetaNetX (MNX) and further GEMap on the compounds without structure information. The number of thermodynamic constraints added to the GEMs in terms of the number of metabolites and reactions with estimated $\Delta_f G^\circ$ and $\Delta_r G^\circ$, respectively, can be found in Table C.3.5. Recon 2 v4 has 53.3% of the unique metabolites with estimated $\Delta_f G^\circ$, resulting in approximately 44.3% the metabolic reactions with thermodynamic constraints (transport reactions have been excluded from these statistics in both networks but their numbers are summarized in Table C.3.5). For iMM1415, about 57% of unique metabolites have estimated $\Delta_f G^\circ$, resulting in approximately 47% of the metabolic reactions in the network (excluding transport reactions) with thermodynamic constraints. These results are summarized in Figure 3.11.

For the case of the mouse model, we further compare the thermodynamics constraints coverage of iMM1415 using the described pipeline (iMM1415) with the one obtained from using the initially released version of the model by relying only on the KEGG compound annotations that have been requested to the authors of this GEM at a time close to publication (*iMM1415). Figure 3.11 shows that a gain of only 57 metabolites with estimated $\Delta_f G^\circ$ translated into a gain of about 2-fold the number of reactions with thermodynamic constraints. This result indicates the importance of

evaluating the highly-connected metabolites (nodes) in the GEM network in order to recover the most reaction thermodynamics constraints. These are metabolites that participate in multiple reactions and in pair with other metabolites and cofactors. Evaluating these highly active network nodes (metabolites) and prioritizing them in the search for chemical structures will improve the reaction thermodynamics coverage.

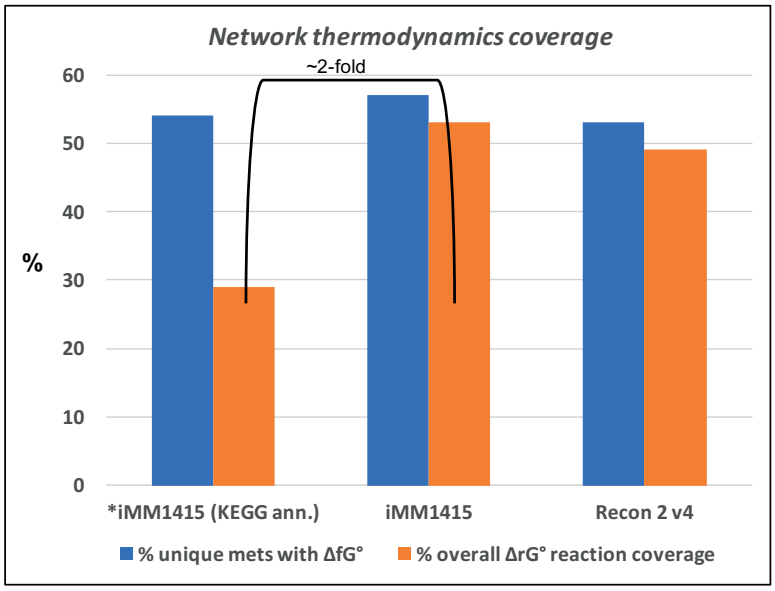


Figure 3.11 Assessment of network thermodynamics constraints coverage for Recon 2 v4 and iMM1415 GEMs mapped with a combination of MetNetX (MNX) and GEMap. *iMM1415 model is shown for comparison with iMM1415 using only the KEGG annotations provided by the authors at the time of publication.

3.4 Conclusion

We have successfully established a pipeline (DRAMA) that facilitates thermodynamic parameter curation and data integration into GEMs. The flexibility of the system doesn't require the use of a unique identifier, although one could be implemented. There are several advantages to the use of such automatized pipeline for searching databases. With these web services, users can implement automated searches at different stages of model analysis or GEM reconstructions. The whole procedure is quite stable, requiring practically no maintenance of the request functions; contrary to database dump files provided in SDF, html, sql, txt, and others formats that require update and more or less

parsing. Modifications and reassessment would only be required if there is a major restructuring of the database web page or if the web service protocol and accesses are modified. Finally, these services provide the most up to date database contents that can be accessed in real time, free of cost and without licensing requirements.

We have demonstrated GEMap to be an apt tool for mapping compound names to external identifiers of several databases (with a mix of local databases and web services requests). Its usefulness, as in any workflow that tries to match nomenclature, is limited by how representative the search name is with respect to the real compound name. When searching external identifiers by compound names, the accuracy of the search result depends on the name of the compound following the rules of chemical nomenclature as used in most databases. For querying databases there are some algorithms for fuzzy string matching that can be applied, such as, the edit distance, Levenshtein distance, and its variants (173). The search through the use of web services provides more flexibility. The search name does not need to be a perfect match to the compound synonyms within the online database. Even though online requests through the web services allow for more flexibility in the compound name structure, all relevant chemical substrings in the name should be present to guaranty a match. Results will be harder to retrieve when abbreviations are used and other string tags (such as formulas or cell compartment) are added to the name. In particular, for chemical names, any number near a charge sign or {R, S, L, D} letters, among others, may lead to different isomers of the same compound. For applications where it is important to differentiate protonation or stereochemistry for the identification of compounds, more advanced string matching tools would need to be applied. This pipeline also heavily depends on the accuracy of the cross-referencing among databases. Despite some discrepancies found regarding structural representation of the compounds, a study performed on compound database cross-referencing has shown that a cross-referencing accuracy of 93% from ChEBI to HMDB and 82% from HMDB to ChEBI (174), two of the seven databases in use. Since the study was performed in 2012 and there is continuous updated of these and other databases, this number may be currently higher. Nevertheless, the flexible structure of the pipeline allows for the user to choose to do processing of compounds primarily with other resources for acquiring external database identifiers as exemplified in section 3.3.3.

The name search feature of GEMap was primarily developed to quickly draft a mapping of compound names in fluxomics and metabolomics datasets without any other available identifier, as well as to annotate GEMs without external identifiers. Until recently, many models were being transferred without this information. Currently, there has been a great deal of effort in annotating some of the older models for public use. In particular, BiGG database has recently made available several older models, including Recon 1 and the iMM1415, with annotated SBML formats.

We also showed the use of GEMap for compound thermodynamics curation in both iMM1415 and Recon 2 v4. These pipelines are independent of the search by compound name and automatically retrieve all compound information by looking up the external identifiers in each respective database. However, currently, the number of compounds that will have an estimation for $\Delta_f G^\circ$ is limited by the presence of R groups in the structure. Since the composition of these groups is unknown, GCM cannot estimate free energy contributions for them. In order to improve the results, we are currently looking into Chemaxon Standardize (161) function to remove these R groups from the molfiles for the computations. Removing these R groups from the compounds should not constitute a problem since our pipeline ensured that reactions with thermodynamics constraints are properly balanced and, as such, these groups should appear unchanged on both sides of products and reactants.

The DRAMA pipeline described in this chapter and applied to the human metabolic model Recon 2 v4 allows for semi-automatization of standard procedures. It improves data management, model annotation, data integration into the GEM and, most importantly, supports reproducibility. This pipeline integrating all stages of model assessment for proper thermodynamics curation and data management sets the standard operating procedure to apply to any GEM with the purpose of performing TFA analysis and for consistently generating condition-specific metabolic network reductions of any GEM, which will be the topic of the next chapter.

Chapter 4: From Human GEMs to consistently derived reduced human metabolic networks

4.1 Introduction

4.1.1 Cancer studies with genome scale models

Recently there has been an increasing interest in the study of cancer by focusing on tumor metabolism with the purpose of identifying both biomarkers and potential drug targets that will specifically impair cancer cell survival or even induce apoptosis of tumor cells. This interest has arisen from the many important discoveries on tumor related metabolic reprogramming, which has become a Hallmark of cancer (175-178). With human GEMs becoming increasingly available (6, 10, 11, 138, 139) for -omics integration, several studies have been systematically performed to identify key features in the metabolic reprogramming of different cancers. Metabolic models are able to contextualize the different types of data (fluxomics, metabolomics, transcriptomics, and proteomics) such that cancer-specific metabolic pathways emerge from the topology of the network.

One of the Hallmarks of cancer is the Warburg effect, which was first observed by Otto Warburg in 1924 (179). The Warburg effect in cancer cells is characterized by aerobic glycolysis in the presence of abundant oxygen, high glycolytic rates, and accumulation of byproduct formation (lactate). Glycolytic enzymes and glucose transporters are found to be over-expressed or deregulated in tumors in support of the observed high glycolytic rate, while in parallel there is evidence of decreased pyruvate transport into the mitochondria, supporting the idea that cancer cells survive without respiration. It is this high glycolytic effect that is the basis of 18-FDG-PET imaging techniques for cancer diagnostic and tumor visualization.

This overflow metabolism correlates with the requirement for upregulation of bioenergetics and biosynthetic pathways to sustain the rapid cell proliferation observed in tumors. Glycolysis provides the metabolic intermediates to support these pathways and, in parallel, the production of NAD+ from the conversion of pyruvate to lactate

continuously supports glycolysis and citric acid cycle reactions. Higher production of glycolytic and citric acid cycle intermediates maintains the pool of NADHP for reducing power that supports biosynthetic pathways. All the metabolic characteristics of the Warburg effect are well summarized and described in these two reviews (176, 177).

The Warburg effect based on the existence of a lactate overflow metabolism under high oxygenated conditions is not a metabolic phenotype uniquely characteristic of disease states, such as in cancer metabolic reprogramming. This metabolic phenotype is also characteristic of normal proliferating cells, as it has been observed for proliferating mouse fibroblasts (180), mitogen-stimulated normal human lymphocytes (181), mouse lymphocytes (182), and rat thymocytes (183). Indeed, a picture seems to emerge where all proliferating cells have high glycolytic rates to sustain their demands, which rather sets the overflow metabolism as a common feature of all growing tissues (184).

However, cancer cells present very diverse metabolic phenotypes that the Warburg effect cannot fully explain. In fact, the extent and presence of the Warburg effect has been observed to vary in different cancers (185). There is a multiplicity of perturbations in the basis of cancer cell metabolic reprogramming heterogeneity. Mutations in oncogenes and on the encoding of tumor suppressor genes, as well as perturbations in cell signaling, can deregulate metabolic enzymes and their pathways and drive metabolic reprogramming in cancer. Differentially expressed enzymes in cancer cells affect nutrient utilization and these overall metabolic modifications can drive tumor proliferation by presenting metabolic phenotypes that diverge from the typical glycolytic fermentation of the Warburg effect and towards a more active citric acid cycle and oxidative phosphorylation (186). Studies in glioma, hepatoma and cancer cell lines have shown that different cancer types actively use oxidative phosphorylation to fulfill energy requirements (187-190). A higher rate of amino acid uptake has been observed in cancer cells under different conditions (191, 192), indicating that carbon sources other than glucose are able to fuel biosynthetic pathways. These effects of differential nutrient uptake in cancer have been observed in culture by performing nutrient replacement or starvation studies (192-196). Amino acids that are not essential in humans, i.e., the cell has metabolic pathways to support their *de novo* synthesis, can thus become condition-specific essential in cancer cells. To add further

complexity to this picture of metabolic plasticity and diversity observed in different cancers to maintain their proliferation and survival, there are studies that show that cancer cells can revert from a Warburg phenotype to an oxidative respiration based on nutrient availability and surrounding conditions (197).

It is this diversity and metabolic plasticity that provides cancer cells with the ability to adapt and proliferate in different and aggressive tissue conditions. Amidst the metabolic heterogeneity observed, the fundamental characteristic of different cancers is that they all evolved from their original tissues to enhance metabolic biosynthetic pathways that support rapid growth and cell proliferation. This observation is in the basis of different studies where a flux solution for the network is determined using FBA with growth maximization as a cellular objective (125, 126, 198). Other objective functions, such as optimization of energy (ATP) production (199) and maximization of redox (NADPH) potential (8) have also been used. These works rely on the computation of one optimal solution of the flux vector in a multiplicity of solutions constrained to the same feasible space that fulfill the same objective.

Recent studies used different levels of data integration into GEMs to analyze cancer metabolic reprogramming. One study integrates transcriptomics data to remove pathways associated with low metabolic enzyme expression levels and used FBA to predict drug targets that inhibit cancer cell proliferation (9). In other works, algorithms for building normal and cancer tissue-specific models based on transcriptomics and proteomics data were developed (INIT (18) and mCADRE (200)) and the resulting networks were used to identify frequent metabolites and pathways occurring in different cancer tissues (18, 200). Other algorithms have been developed to predict metabolic fluxes based on tissue specific expression data (GIMME (13) and iMAT (15)). iMAT was used to investigate metabolic differences in HepG2 cancer cell lines expressing different levels of oncogene p53 (201). All the works inferring network properties from transcriptomics are based on the assumption that metabolic enzyme levels correlate with gene expression levels, independently of any downstream regulatory effects that may be taking place. Despite the considerable amount of cancer studies using human GEMs with -omics integration that contributed to the elucidation of pathways essential for cancer cell proliferation, their intrinsic mechanisms and how

the flux through the reactions in these pathways is regulated is still not fully understood, as it relies on a combination of oncogene expression, pressure from the microenvironment, and local nutrient availability (178, 202). Much effort is still needed to bring this understanding to a level where efficacious therapies that target metabolism can be developed and put into the clinic.

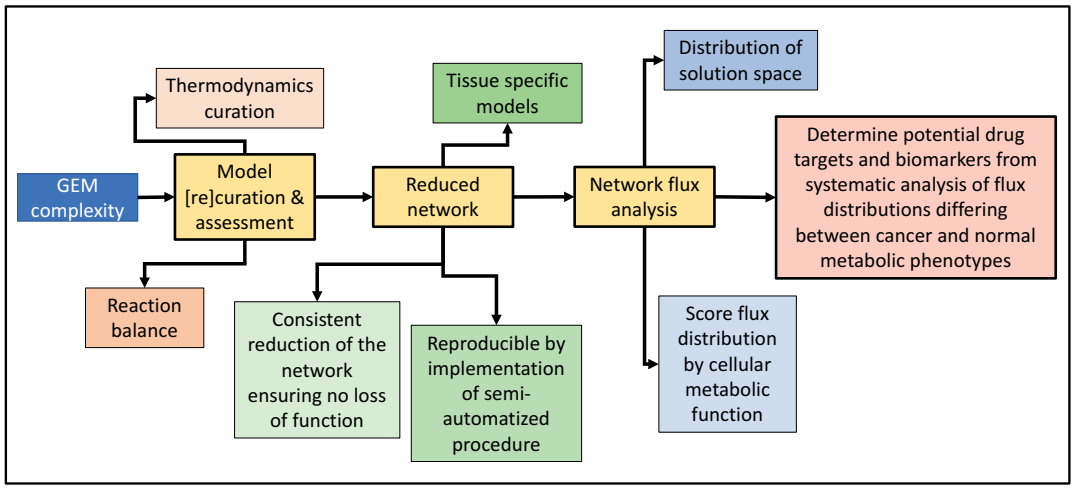


Figure 4.1 Complete pipeline for processing and reducing a human GEM for the study of cancer metabolic rewiring and heterogeneity.

We propose a pipeline (Figure 4.1) for generating cancer tissue-specific models that are focused on the subnetworks of interest for the study of the different cancer types. Starting from a human GEM we apply the pipeline described in Chapter 3 as a standard operating procedure to assess and [re]-curate the GEM with missing annotations and parameters. This process involves identifying metabolites and mapping them to external database identifiers with the purpose of balancing the network, integrating thermodynamics constraints for the reactions and matching the compounds to the available data. Extracellular fluxomics and intracellular metabolomics data pertaining to the cancer types and conditions that will be studied are also processed within the described pipeline. The next step in this pipeline is to reduce the network in a consistent and semi-automatized way to generate tissue-specific models that represent the cancer physiology of interest. which is the focus of this chapter. The last part of the pipeline consists in the analysis of the generated physiologies through systematic analysis of flux distributions given the applied condition-specific data and

thermodynamics constraints, with the purpose of determining drug targets and biomarkers in cancer, as well as to determine the differences between normal and cancer physiology. The analysis as proposed by this pipeline differs from the typical FBA computation of one solution for imposed metabolic objectives in the sense that it focus on the analysis of the complete feasible space of solutions, which can then be classified by a score in terms of maximal growth, ATP production, redox potential, and so on (see discussion in Appendix A.4.1 for an example with a reduced model obtained from Recon 1). Nevertheless, the focus of this chapter is the establishment of the steps for the generation of a reduced human metabolic model from a GEM, and not the subsequent analysis.

4.1.2 Consistent decrease of GEM complexity tailored to system under study

Analyzing GEMs with networks of the size of human GEMs (such as Recon 2 v4 and HMR) can become a cumbersome task, especially when the intracellular metabolomics and extracellular fluxomics data available to integrate into the model is scarce. Typically, measured metabolites are the ones participating in reactions of the central carbon pathways and important lipids, leaving most of a GEM network uncovered. Furthermore, these networks have a lot of flexibility due to the high number of unknown reaction directionalities, and sampling methodologies for the extraction of useful information within the feasible space of solutions become very time-consuming. The most practical analysis with GEMs is thus the use of FBA with optimization towards a cellular metabolic objective, such as maximization of growth yield, or ATP maintenance, among others. As mentioned above, despite its usefulness, this approach provides one solution for the directionalities and magnitude of the fluxes in the network that is not unique and certain key aspects may be overlooked, such as the account of all physiological possibilities.

Reducing the GEM for the study of the organism and condition of interest is common practice in this field to overcome complexity and yield more insightful analysis. In fact, in the case of Recon2 v4, FVA and TFVA have shown that about ~30% of the reactions will be blocked in the network, as shown before in Table 3.4. Commonly used approaches to reduce the GEM complexity make use of proteomics and mRNA

expression data to further constrain or reduce the size of the network (13-15, 18, 200). Other approaches focus on selecting subnetworks from the GEM pertaining only to the parts of the metabolism of interest, as well as pathways that fulfill metabolic requirements relevant for the organism being studied (8, 203, 204). In this work, we use redGEM (205, 206), a workflow developed in our lab (details in section 4.2.3), to systematically reduce GEMs into phenotypically-driven core networks in a semi-automatized and consistent way that ensures reproducibility. Thought this systematic reduction procedure that allows the subsequent shrinking and expansion of the network, the model is tailored to the physiology of interest and it becomes a tool for the comparative analysis of different phenotypes.

4.2 Materials and Methods

4.2.1 Recon 2 preprocessing

Recon 2 v4 network (dated 11.05.2015) was downloaded from the Virtual Metabolic Human website (http://vmh.uni.lu/#downloadview). The model was processed with the pipeline described in Chapter 3 for metabolite identification and retrieving compound structural information. Metabolite structural information was used for balancing reactions and for reaction thermodynamic curation to determine reaction directionality as described in Chapter 3.

4.2.2 Data & parameter integration into Recon 2

We collected datasets with extracellular fluxomics and intracellular metabolomics data from multiple sources spanning a range of human normal and cancer cell types or tissues. Extracellular fluxomics datasets contain measured uptake and secretion rates, whereas intracellular metabolomics datasets contain measured intracellular concentrations. Metabolites quantified are mostly amino acids, glucose, lactate and other compounds pertaining to pathways of the central carbon metabolism.

A summary of the types of data collected per tissue type for normal (NH) and cancer (CH) is displayed in Table 4.1 and Table 4.2, respectively. In total, there are 11 datasets for the normal phenotype (NH5 was used uniquely for data on cell growth rate), and 69 datasets for the cancer phenotype (CH5, CH6, CH72 and CH73 were used uniquely for data on cell growth rate). Detailed information for each normal and cancer dataset can be found in Appendix C: see Table C.4.1 for a complete dataset description and experimental sources; see Table C.4.2 for cell growth rate measurements associated with each dataset; and see Table C.4.3 for the conversion factors used to get fluxes in standard units of mol/h/gDW and concentrations in mol/L. In general, when the cell dry weight is not known for the cancer cell lines for which we have flux measurements, we assume the cell dry weight of the HeLa cell (~400pg). This cell weight is close to the average value estimated for different mammalian cells at different conditions in (207).

We use the DRAMA pipeline described in section 3.2 of Chapter 3 and summarized in Figure 4.2 for preprocessing Recon 2 v4 and the datasets, and for systematically integrating the data into the model. Metabolomics and fluxomics data can be aggregated per tissue type or cell type or phenotype as shown in Table 4.3 in order to simulate the metabolic physiology of interest. To generate the first reduced human metabolic network with redGEM that comprises both the normal and cancer physiologies (*RedHuman*), we integrated the All Human dataset in Recon 2 v4 model (Table 4.3) that merges the 84 fluxomics and metabolomics sets for cancer and normal conditions.

The basal value of ATP maintenance for mammalian cells was assumed to be the same as the one measured for the mouse fibroblasts LS cell line 1.77 mmol/h/gDW (converted from 1.7e-11 molATP/cell/day) (208).

Since cancer specific cellular biomass quantification is unavailable, we chose to keep at this stage the biomass reaction included in Recon 2 v4 GEM.

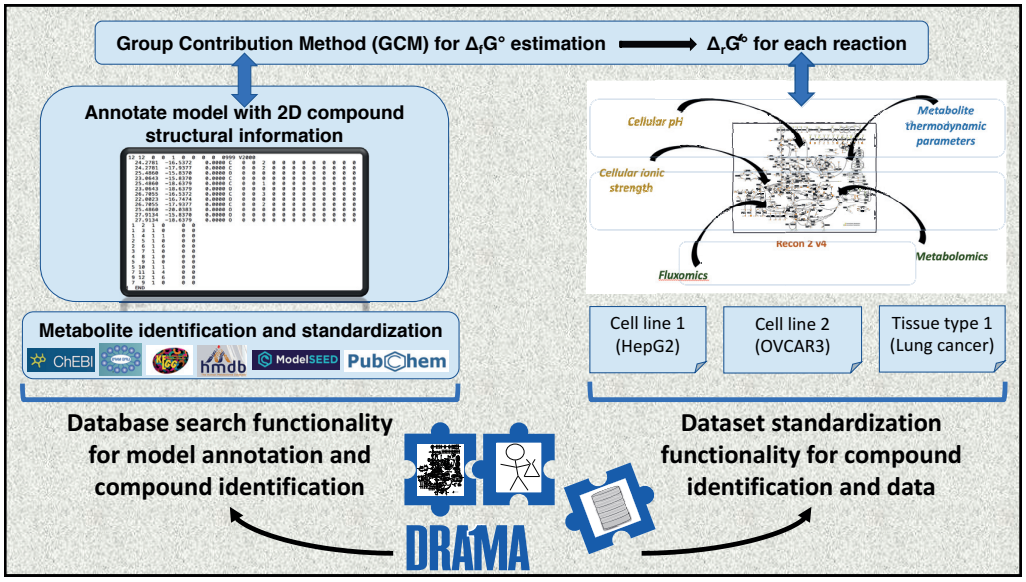


Figure 4.2 Overview of DRAMA full pipeline applied to Recon 2 v4 GEM. The GEM is mapped with GEMap and MetanetX (see section 3.2.2 for more details), followed by automatic curation of metabolite structural information (see section 3.2.3 for more details), which is used for estimation of $\Delta_r G'^{\circ}$. In parallel, the fluxomics and metabolomics datasets for healthy human and cancer cells are preprocessed with GEMap and matched to the metabolites in the Recon 2 v4.

Table 4.1 Summary of human normal (NH) datasets types collected per tissue. The * indicates datasets used just for growth rate estimation.

Normal Human	Tissue Type	Extracellular Fluxes	Intracellular concentrations
NH1	kidney	X	
NH2	kidney		X
NH3	liver		X
NH4	lung	X	X
NH5*	liver		
NH6	colon		X
NH7	lung		X
NH8	prostate		X
NH9	stomach		X
NH10	prostate		X
NH11	lung		X
NH12	lung		X

**** Used just for growth rates***

Table 4.2 Summary of human cancer (CH) datasets types collected per tissue. The * indicates datasets used just for growth rate estimation.

Cancer Human	Tissue Type	Extracellular Fluxes	Intracellular concentrations	Cancer Human	Tissue Type	Extracellular Fluxes	Intracellular concentrations
CH1	lung	X		***CH38***	leukemia	X	X
CH2	lung	X		***CH39***	ovary	X	X
CH3	colon		X	***CH40***	lung	X	X
CH4	cns	X		***CH41***	lung	X	X
CH5*	liver			***CH42***	lung	X	X
CH6*	liver			***CH43***	lung	X	X
CH7	liver	X	X	***CH44***	lung	X	X
CH8	liver		X	***CH45***	ovary	X	X
CH9	lung		X	***CH46***	ovary	X	X
CH10	kidney	X	X	***CH47***	ovary	X	X
CH11	kidney	X	X	***CH48***	ovary	X	X
CH12	lung	X	X	***CH49***	prostate	X	X
CH13	kidney	X	X	***CH50***	leukemia	X	X
CH14	breast	X	X	***CH51***	kidney	X	X
CH15	kidney	X	X	***CH52***	cns	X	X
CH16	leukemia	X	X	***CH53***	cns	X	X
CH17	colon	X	X	***CH54***	cns	X	X
CH18	prostate	X	X	***CH55***	skin	X	X
CH19	lung	X	X	***CH56***	skin	X	X
CH20	colon	X	X	***CH57***	skin	X	X
CH21	colon	X	X	***CH58***	ovary	X	X
CH22	colon	X	X	***CH59***	kidney	X	X
CH23	leukemia	X	X	***CH60***	cns	X	X
CH24	lung	X	X	***CH61***	cns	X	X
CH25	lung	X	X	***CH62***	leukemia	X	X
CH26	breast	X	X	***CH63***	colon	X	X
CH27	colon	X	X	***CH64***	breast	X	X
CH28	ovary	X	X	***CH65***	kidney	X	X
CH29	leukemia	X	X	***CH66***	cns	X	X
CH30	colon	X	X	***CH67***	skin	X	X
CH31	skin	X	X	***CH68***	skin	X	X
CH32	skin	X	X	***CH69***	kidney	X	X
CH33	lung	X	X	***CH70***	lymphatic tissues/ blood cells	X	
CH34	breast	X	X	***CH71***	prostate		X
CH35	breast	X	X	***CH72****	stomach		
CH36	skin	X	X	***CH73****	prostate		
CH37	breast	X	X				

Table 4.3 List of aggregated datasets to build different metabolic tissue phenotypes for model reduction and analysis.

Tissue Type	# of sets	Data Sets ID
Lung	10	*CH12,CH19,CH24,CH25,CH33,CH40,CH41,CH42,CH43,CH44*
Skin	8	*CH31,CH32,CH36,CH55,CH56,CH57,CH67,CH68*
Kidney	8	*CH10,CH11,CH13,CH15,CH51,CH59,CH65,CH69*
Ovary	7	*CH28,CH39,CH45,CH46,CH47,CH48,CH58*
Colon	7	*CH17,CH20,CH21,CH22,CH27,CH30,CH63*
Breast	6	*CH14,CH26,CH34,CH35,CH37,CH64*
CNS	6	*CH52,CH53,CH54,CH60,CH61,CH66*
Leukemia	6	*CH16,CH23,CH29,CH38,CH50,CH62*
Prostate	2	*CH18,CH49*
All Lung Cancer	13	*CH1,CH2,CH9,CH12,CH19,CH24,CH25,CH33,CH40, CH41,CH42,CH43,CH44*
All Lung Normal	4	*NH4,NH7,NH11,NH12*
All Cancer	73	*All CH# sets*
All Normal	12	*All NH# sets*
All Human	84	*All NH# and CH# sets*

4.2.3 Reduction of GEMs to data-driven networks

We generated *RedHuman*, a metabolic network reduced from Recon 2 v4 containing the combined extracellular fluxomics and intracellular metabolomics for all cancer and normal conditions (AllHuman). For the reduction procedure, we used redGEM (205), a workflow developed in our lab, to reduce GEMs into phenotypically data-driven networks. Contrary to many early attempts at producing reduced models, this workflow has the ability to expand and shrink the network in a systematic and consistent way that allows for tailoring the model to answer the relevant physiological questions, while being performed in a semi-automatic and systematic way that is consistent with predefined criteria and hence reproducible. The redGEM workflow is summarized in Figure 4.3 and has four main stages.

Briefly, in stage 1 we preprocess the GEM by selecting the core subsystems and reactions (core network) of interest, which will be the main focus with respect to the phenotype/condition being studied, and phenotypic data integration, such as extracellular fluxomics and intracellular metabolomics. Integrating data during the reduction procedure ensures that the key network properties of the physiology of interest will emerge and be kept.

On stage 2 we expand the network via graph search by recovering from the GEM the pathways/reactions, which are selected based on the number of reaction steps connecting the core metabolites being below a user defined threshold. Although the reactions in the selected core subsystems are initially preserved, some reactions will not carry flux and will thus be removed from the network.

On stage 3 we use lumpGEM (206), a MILP (mixed-integer linear programming) procedure developed to produce lumped reactions by finding the shortest pathways connecting metabolites in the core network to their final biomass building blocks. The process takes into account the original cell composition of the organism integrated into the GEM (biomass reaction) and the existence of alternative biosynthetic pathways for the production of the same biomass building block and generates alternative lumped reactions. This procedure ensures that mass balance is preserved and that the reactions are thermodynamically feasible and carry flux.

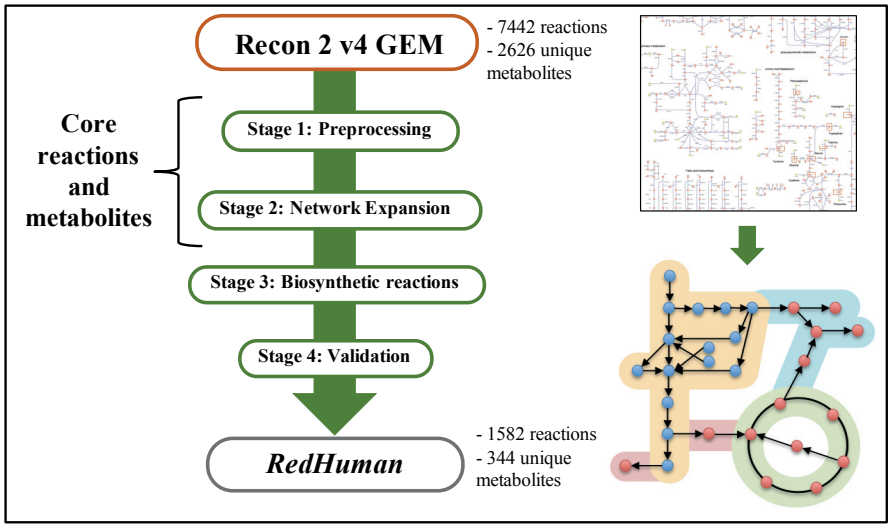


Figure 4.3 Summary of redGEM workflow for the systematic, semi-automatized reduction of GEMs into data centric networks.

In the last stage, we validate the reduction by checking that the GEM metabolic capabilities are maintained, such as maximum growth yield and the performance of metabolic tasks pertinent to the physiology of the organism represented by the GEM, and now by the reduced model.

4.3 Results

4.3.1 Generating data-driven reduced models

As illustrated in Figure 4.4, redGEM can be used to systematically reduce models for different physiologies given the data presented in Table 4.3. In this section, we present the reduction procedure of Recon 2 v4 with all normal and cancer data (All Human) integrated for illustration. The resulting network is referred to as *RedHuman* and is able to represent both cancer and normal physiologies once the All Human data is removed and the All Cancer and All Normal datasets (Table 4.3) are individually integrated into the *RedHuman* model. This network can be used to study the differences between normal metabolism and metabolic reprogramming associated with cancer.

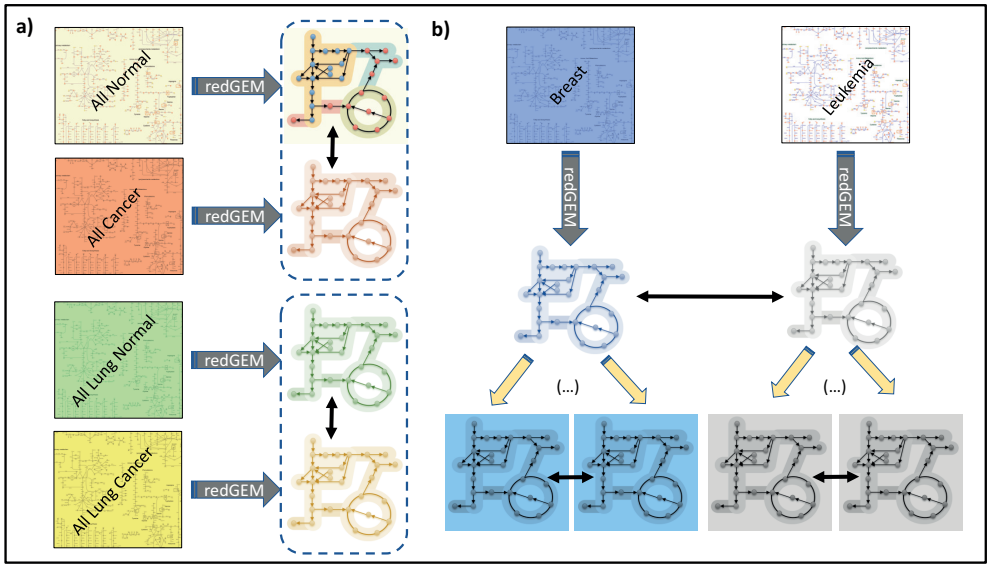


Figure 4.4 Illustration of the pipeline for generation of reduced models with redGEM. Different reduced models can be generated with data pertaining to different physiologies. After reduction, each cancer tissue specific model (such as breast and leukemia) can be repopulated with the fluxomics and metabolomics of each individual cell line of the same tissue to study the underlying variability of cancer metabolic reprogramming within the tissue.

4.3.1.1 Assessment of dataset variability

Before integrating the data, we assessed the variability across the datasets to ensure they can reflect physiological differences. The datasets showed evidence of metabolomics and fluxomics variability that can be explored in the analysis of the

different phenotypes. Survey of metabolomics from all NCI60 cancer cell lines included in the cancer datasets showed an overall variability in the level of intracellular metabolites across the 60 cell lines (Figure B.4.1), which is also observed, in smaller extent, when the levels of metabolites for different cell lines from the same tissue are compared (Figure B.4.2 for NCI60 prostate and Figure B.4.3 for NCI60 leukemia, for instance). Interestingly, prostate tissue with only two cell lines (Figure B.4.2) presents more variability in the measured intracellular levels than leukemia with 6 cell lines (Figure B.4.3). See Figure B.4.4 for All Normal vs. All Cancer comparison.

4.3.1.2 Minimal media and assessment of extracellular reactions (uptakes/secretions)

The procedure of data integration involved several steps to ensure the models are functional using minimal assumptions. To further constrain the physiology, we started by blocking the uptakes of metabolites that are not typically transported across the cell membrane. These metabolites are molecules that contain moieties, such as coenzyme A (CoA), acyl-carrier protein (ACP) or phosphate groups, that are not transportable through the membrane by simple diffusion (see the list in Table C.4.4). These moieties require specific transport mechanisms, which if not identified in the model, should not be allowed. On the other hand, even if the transport existed it would allow only for a very low maximum uptake rate and the cell would still be required to produce these moieties. For instance, the *de novo* synthesis of CoA is a conserved pathway across organisms (209).

Apart from the network size, the complexity of human metabolism lies on its redundancy, i.e., the possibility of cell growth on alternative carbon sources. These alternatives translate into tissue- and condition-specific changes in metabolism, and thus we should ensure these extracellular reactions take part in the landscape of possible physiologies in our analysis even when no extracellular fluxomics data is available. These alternatives are found through *in-silico* minimal media analysis, where growth in the GEM is explored by selecting minimal sets of nutrient uptakes. The extracellular reactions for these nutrients that are essential for cellular growth and for which there were no extracellular flux measurements, are then forced to be opened with a general high upper bound flux constraint. Extracellular reactions for other metabolites that are commonly found in the minimal media used in mammalian cell

culture and in serum (210, 211) are also ensured to be unblocked in the model. Table C.4.5 has the complete list of selected metabolites for which extracellular fluxes are checked independently of data, but their flux constraints are overwritten in case measured uptakes are available in the dataset.

4.3.1.3 GEM assessment with data for specific physiology

Before starting the reduction procedure and after taking care of the steps above, Recon 2 v4 was tested by integrating the datasets presented in Table 4.3. This is done to ensure the GEM is functional with the data prior to reduction. The data for the different physiologies was added by stages and the models were evaluated for maximal growth: models with data should be able to produce biomass, but also to sustain growth yields similar to the maximum growth measured for the cell lines included in the different physiologies. We note that when testing the GEM, the growth yield is often much higher than the one obtained from the measurements when other uptakes for which there is no extracellular flux data remain unconstrained.

The fluxomics data containing uptake and secretion rates for each physiology in Table 4.3 was initially added without the intracellular metabolomics data and FBA was used to compute maximal growth yield under the specified conditions. Uptake and secretion rates smaller than 1e-7 were excluded from the dataset. Colon, lung, ovary and All Lung Cancer models failed to grow at this stage. Since in FBA no thermodynamic constraints are imposed, failure to grow is solely an effect of both the imposed extracellular flux constraints and the mass-balances of the network. Further analysis for uncovering minimal sets of data that are problematic for colon, lung, ovary and All Lung Cancer models revealed that the extracellular fluxes of D-sorbitol and the vitamin Thiamine were the problematic data points. For these four models in particular, both Thiamine and D-sorbitol were measured secretions. Thiamine is a required nutrient in cell culture and essential to humans. It is present in the minimal media used for the NCI60 cell lines and it is a known unstable molecule in culture as indicated in (http://www.sigmaaldrich.com/life-science/cell-culture/learning-center/media-expert/thiamine.html). This indicates that the experimental measurement of Thiamine as secretion could be an error. The problem with D-Sorbitol was related with an unsustainable forced lower bound of its secretion. With these extracellular uptakes

corrected all four models resumed the desired growth yield with FBA analysis and it was maintained after adding thermodynamic constraints for the reactions and concentration measurements.

After applying thermodynamics constraints and adding intracellular concentrations to the models, we found that all models required at some extent relaxation of the lower bound of D-sorbitol secretion. Breast, cns, kidney, leukemia, prostate, skin and All Cancer models also required relaxation of either one or a pair of uptake/secretions of creatinine, adenosine, inosine or intracellular concentration of glycine. The need for relaxing data constraints is a typical occurrence when integrating data into GEMs, and it is a direct consequence of measurement uncertainties and/or underlying model assumptions.

4.3.1.4 Selection of core subsystems for reduction

The next step is to select the core subsystems of interest to be kept in the *RedHuman* core network. The list of core subsystems selected comprises: *Glycolysis/gluconeogenesis; Citric acid cycle; Glutamate metabolism; Oxidative phosphorylation; Pentose phosphate pathway; Pyruvate metabolism; Glutathione metabolism; Glycine, serine, alanine and threonine metabolism; Arginine and proline metabolism; Cysteine metabolism; CoA synthesis; ROS detoxification; Cholesterol metabolism; Fatty acid synthesis*.

This list contains all the main subsystems present in central carbon metabolism. These are the subsystems of interest to study the overflow metabolism characteristic of the Warburg effect, a Hallmark of cancer metabolic reprogramming, where the key features are a high glycolytic activity leading to lactate byproduct formation with (sometimes) deregulation of citric acid cycle. There are also other subsystems included that have been observed to play an active role in cancer proliferation and heterogeneity.

Glutamate metabolism and Citric acid cycle subsystems connect the glutaminolysis pathway, which with increased glutamine uptake rate plays an important role in cancer metabolism, fueling parts of the citric acid cycle to sustain biosynthetic pathways (212, 213). Glutamine is also involved in the production of glutathione to counteract and detox the high levels of reactive oxygen species (ROS)

formed as a result of increased metabolic activity within the cell (214). There is however heterogeneity in cancer metabolism with respect to glutamine being uptaken or synthesized, which has been observed in breast cancer occurring in different tumor regions (215, 216) and as well in different regions of the same NSCLC tumor that differ in their microenvironment (202).

Non-essential amino acids glycine and serine are precursors for the synthesis of proteins, lipids, nucleotides and glutathione metabolism, and their biosynthesis pathways and are often found to be upregulated in cancer (217, 218). Serine is also a precursor for the biosynthesis of other non-essential amino acids and participates in the reactions leading to the production of sphingolipids, phospholipids and nucleotide synthesis, which have been implicated in the support of cancer cell proliferation (219). In a screening of 60 different cancer cell lines (NCI60 cell lines in (217), which are part of our datasets), glycine consumption was found to correlate with rapidly proliferating cells, which suggests that its endogenous biosynthesis is not sufficient to support the growth of fast replicating cells (217). Breast cancer studies have observed a coupling between serine secretion and glutamine uptake, which has also been identified as the nitrogen donor for serine biosynthesis (220, 221).

Arginine is considered a semi-essential amino acid in humans despite the existence of a pathway for its endogenous biosynthesis in cells, and it has been widely reported as an essential amino acid for cancer cell proliferation and tumorigenesis (222). Most of the arginine *de novo* synthesis occurs in the kidney, in the proximal renal tubule *via* urea cycle. The cells primary source of arginine is from diet and protein turnover (223). Endogenous arginine biosynthesis is not sufficient to sustain grow of proliferating cells (224). Arginine is involved in the synthesis of proline and glutamate and is also a precursor for the synthesis of polyamines, creatine, nucleotides and nitric oxide (223, 225). Upregulation of these biosynthetic pathways has been shown to promote cancer proliferation, invasion, and metastasis (225, 226). Studies in tumor cells have shown that arginine plays a role in cancer cell proliferation (227-229), which has been further complemented by evidence that arginine starvation can negatively impact tumor growth (228).

The arginosuccinate synthetase reaction (ASS), where citrulline and aspartate are converted into argininosuccinate, is the rate limiting step in arginine synthesis

(223) and the gene coding this enzyme has been observed to be differentially expressed over a wide range of cancer tissue samples (222, 230, 231). Several studies have shown that hepatocellular, renal and prostate carcinomas, as well as some malignant melanomas and pleural mesotheliomas do not express (ASS). For these cancers consumption of extracellular arginine becomes essential, as shown by the effect of arginine degrading enzyme on tumors (232-234).

Lipid metabolism is also observed to be upregulated in cancer. Lipids play a role in the production of cell membranes, which are highly demanded in cancer cells due to their rapid proliferation. They are also involved in signaling pathways necessary for cancer cell survival and progression and in post-translation modification of proteins. The high levels of lipogenesis in cancer cells have been shown to promote cancer cell proliferation and survival (235) and has led to the development of inhibitors targeting enzymes of the fatty acid biosynthesis pathways to suppress cancer cell growth (236-239).

Cholesterol metabolism also plays a role in cancer proliferation. Cholesterol is a precursor of sterols and isoprenoids through the mevalonate pathway, which has been observed to be unregulated in cancer and linked to tumor growth (240, 241). This observation has led to different studies where the effect of statins (a drug used to decrease cholesterol synthesis) on the decrease of cancer cell proliferation was investigated (242-246).

ROS detoxification and glutathione metabolism are key subsystems in cancer due to the balance that cancer cells must achieve to counteract the oxidative stress by producing antioxidants (247). The high levels of reactive oxygen species produced (ROS), although toxic when in high amounts in cells, are known to promote cancer cell survival by causing DNA damage and inducing mutations that promote tumor growth.

The subsystems for fatty acid synthesis and cholesterol metabolism were also included in the core network since metabolic alterations in fatty acid metabolism have been amply reported in recent studies. Lipids are an essential component of cell membranes and cancer cells have been observed to actively synthesize fatty acids to support their proliferation (235). Cholesterol also takes part in the synthesis of cell membranes but is also the precursor for the synthesis of sterols and isoprenoids

through the mevalonate pathway, whose highly-expressed genes have been linked to poor prognosis for breast cancer patients (248) and to cancer progression (249).

4.3.1.5 Reaction thermodynamics coverage per subsystem in the GEM

We have assessed the thermodynamic constraints curation per metabolic reaction (excluding transports) pertaining to different cellular subsystems in Recon 2 v4, which is shown in Figure 4.5 (see Figure B.4.5 for this quantification in terms of number of reactions). Subsystems belonging to the central carbon pathways (glycolysis/gluconeogenesis, citric acid cycle, pyruvate metabolism, oxidative phosphorylation and pentose phosphate pathway) and nucleotide synthesis (purine and pyramidine synthesis, nucleotide sugar metabolism and nucleotide salvage pathway) have 100% of metabolic reactions with thermodynamic constraints. Subsystems consisting on amino acid metabolism have more than 85% metabolic reactions with applied thermodynamic constraints. Other subsystems of interest for cancer cell metabolic reprogramming such as glutathione metabolism, ROS detoxification, cholesterol metabolism, glutamate metabolism, and other non-essential amino acid metabolism have more than 80% of metabolic reactions with thermodynamic constraints.

4.3.2 *RedHuman* assessment & validation

The *RedHuman* network was generated with redGEM with the starting core subsystems presented in section 4.3.1.4 and the All Human data for normal and cancer physiology. The core network expansion in stage 2 of redGEM was performed by connecting the core metabolites among the selected starting subsystems up to three step reactions. The initial human metabolic network Recon 2 v4 had 7442 reactions, of which 2520 were transport reactions (34% of the network), and 2626 unique metabolites. The *RedHuman* network generated (Figure 4.6) contained 2239 reactions, of which 15 are lumped reactions to biomass building blocks and 1415 are transport reactions (63% of the network), and 453 unique metabolites. In the overall, the network reduction resulted in a 5.8-fold and 3.3-fold decrease in the number of unique metabolites and reactions, respectively. Due to the complexity of the transport reactions in human GEMs and their

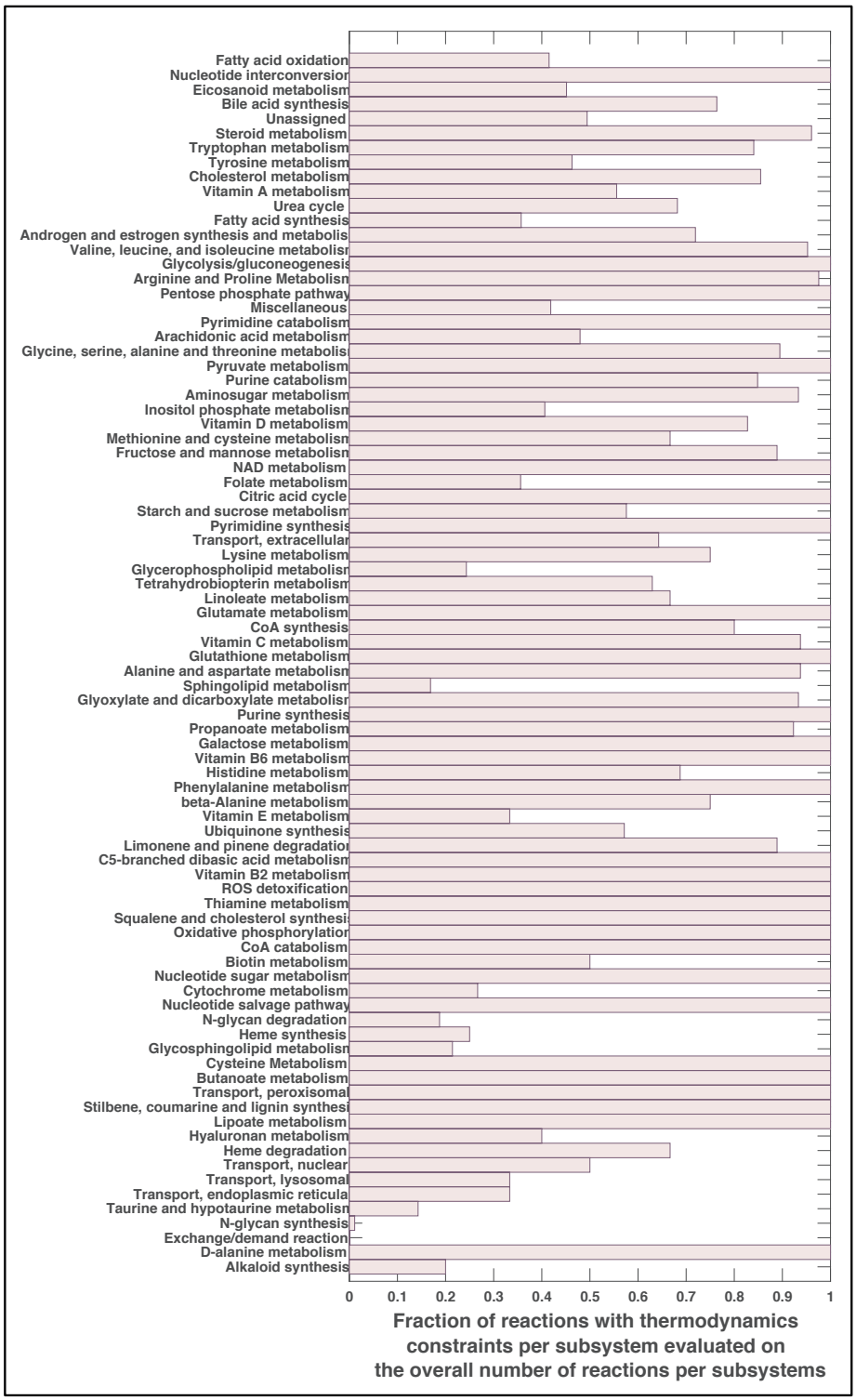


Figure 4.5 Fraction of metabolic reactions per subsystem (excluding transport reactions) in Recon 2 v4 that have reaction thermodynamics constraints.

many forms of carbon sources exchanges, these are chosen to be kept in *RedHuman* and hence the final network is comprised of a large proportion of these (63%). However, we note that *RedHuman* only contains reactions that carry flux given the data and thermodynamics constraints.

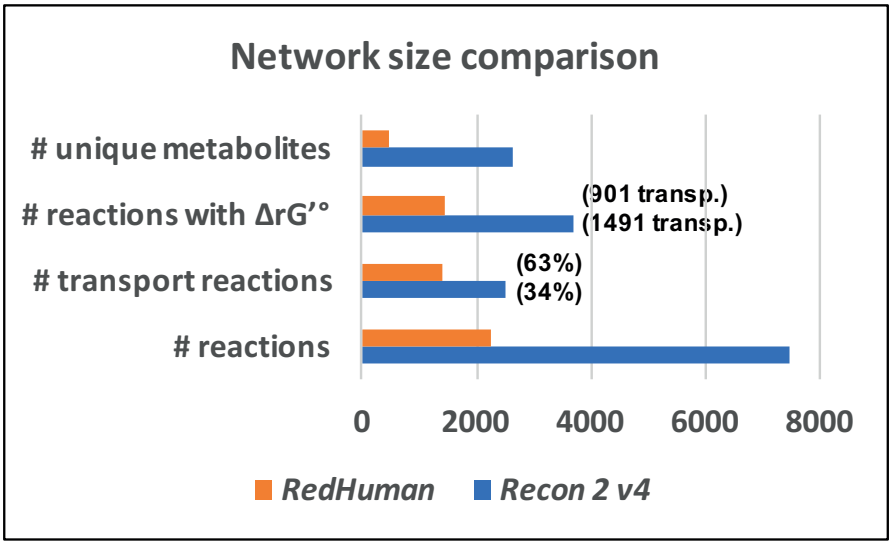


Figure 4.6 Comparison of metabolic network size between Recon 2 v4 and *RedHuman*.

The *RedHuman* has to retain metabolic capabilities that are relevant for the physiology under study. The starting subsystems have been selected as to include parts of the metabolism that may differ depending on cancer types and conditions, such this metabolic reprogramming variability will be captured by the *RedHuman* network. In addition to this it is also important to identify common metabolic tasks, e.g. amino acid biosynthesis pathways for non-essential amino acids and their auxotrophy, which are relevant for the conditions that will be studied, as well as to retain other human metabolic capabilities of the departing GEM. In the overall, all these steps ensure that the resulting *RedHuman* network will have decreased the GEM complexity and its redundancy, while keeping the network flexibility to synthesize the required biomass building blocks.

The network landscape of *RedHuman* is summarized in Figure 4.7 by quantifying the number of reactions per subsystem that remain in *RedHuman* with respect to the initial number in Recon 2 v4. The *RedHuman* network does not include all the

subsystems initially present in Recon 2 v4 (see Table C.4.6 for the list of subsystems not included in *RedHuman*). The *Core Subsystems* are the starting subystems selected at stage 1 of the reduction procedure and the subystems identified by the *Network Expansion* label are the ones that have reactions that connect the initial network defined by the starting subsystems during stage 2 of the reduction procedure. As expected, the starting subsystems have the most coverage on number of reactions kept in the *RedHuman*. We also observe that parts of the metabolism that are closely related to the metabolic functions of the starting subystems emmerge during the network expansion procedure to complement the network funcitonality. This is for instance the case of subsystems that separate anabolic and catabolic counterparts, such as CoA synthesis/CoA catabolism and Fatty Acid Synthesis/Fatty Acid Oxidation. In addition, other subsystems for lipid and nucleotide metabolism emerge during network expansion to complement the central carbon pathways and the routes for production of biomass building blocks. This demonstrates the power of the reduction procedure used as it differs from an 'ad hoc' reduction of the GEM where parts of the metabolism are selected and removed independently of connectivity.

Since amino acid metabolism plays a huge role in cancer associated metabolic rewiring it is important to keep in the *RedHuman* the amino acid biosynthetic capabilities of the GEM. Validation analysis of first attempted reductions of Recon 2 v4 (Figure B.4.6) showed that the amino acid de novo biosynthesis pathways were not completely active in the resulting network. The extracellular reactions allow for auxotrophy of the non-essential amino acids and during reaction lumping to the non-essential amino acids present in the biomass composition the shortest route is then for the cell to immediately uptake those amino acids to go directly into the biomass, hence rendering the amino acid biosynthesis pathways useless. In order to ensure that the amino acid biosynthetic pathways are included in the network as core reactions or through lumped reactions, we block the uptake of the non-essential amino acids during stage 3 of the reduction where the production of biomass building blocks is tested. At the last stage of the reduction procedure, we validate the presence of the non-essential amino acid biosynthetic pathways by testing their synthesis from glucose.

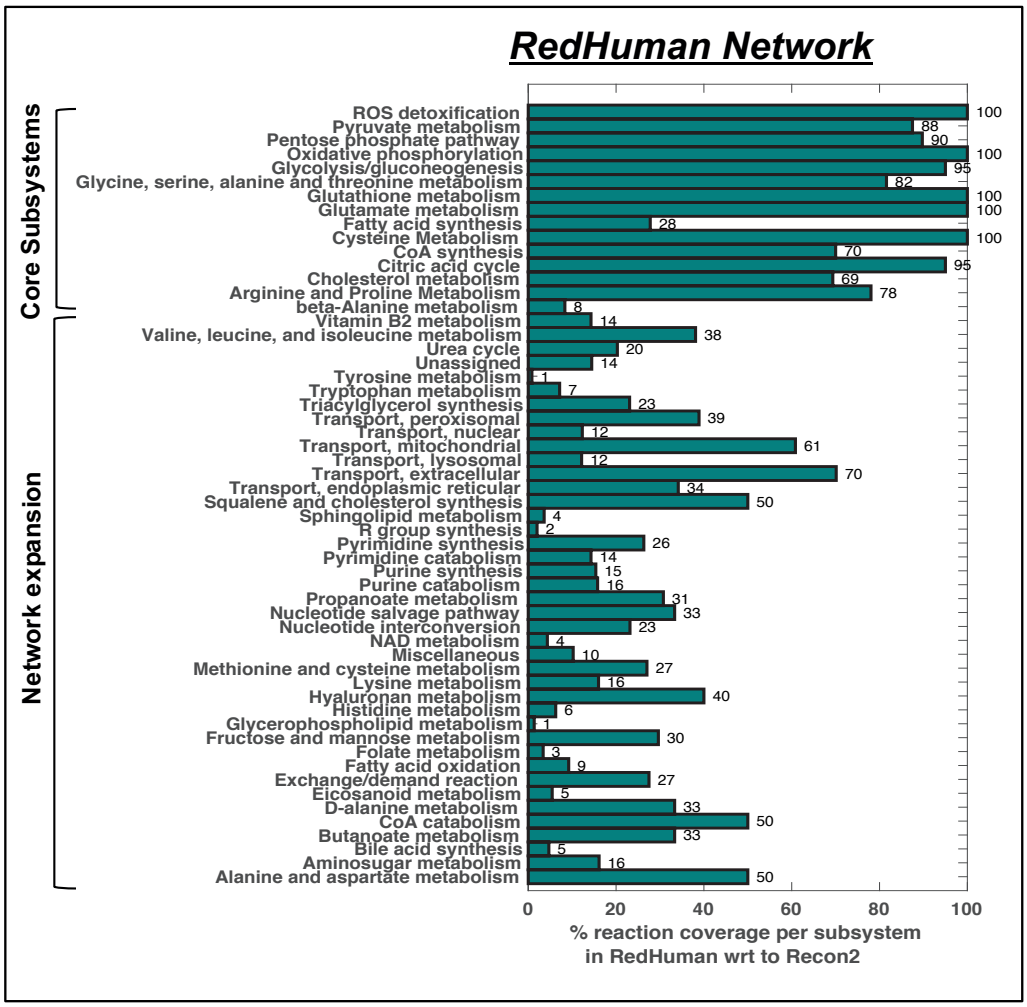


Figure 4.7 Quantification of reactions in the *RedHuman* network with respect to the original Recon 2 v4 network. Quantification is performed by the number of reactions per subystem included *in RedHuman* network with respect to the original number of reactions per subsystems in Recon 2 v4. *Core subsystems* are the subsysystems selected as relevant for the physiology to be studied and they constitute the starting subsystems (along with the subystems containing the extracellular reactions) to initialize the reduction procedure. The subystems identified by the *Network Expansion* label are the ones that have reactions that connect the initial network defined by the core subsystems.

The generated *RedHuman* network has passed on the validation of all metabolic tasks for amino acid biosynthesis, except for L-arginine biosynthesis. We note here that this is not a characteristic of the *RedHuman* construction since it is also observed in the GEM with the different datasets integrated. In fact, this observation is a direct result of the application of reaction thermodynamics to the model (Figure B.4.7). Arginine biosynthesis consists in two reactions Argininosuccinate synthetase (ASS) and

Argininosuccinate lysase (ASL) that convert aspartate and citrulline into L-arginine and fumarate by means of an intermediary product called argininosuccninate. These two reactions are thermodynamically feasible in both directions given a wide range of the intracellular concentrations for their participating substrates and products. However, for the intracellular levels of L-arginine (>1e-7 M) as the ones measured for cancer and normal cells found in our collected dataset (and also widely reported in the literature), the ASL reaction becomes unfeasible in the forward direction, i.e., in the direction of L-arginine production. The observed incapability of synthesizing L-arginine, rather than being a systemic problem, is actually an interesting observation. The observed L-arginine auxotrophy and the downregulation of ASS enzymes in many cancers have propelled the interest on L-arginine starvation as a therapeutic target. However, recently, it has been observed that the resistance of cancer cells treated with long term L-arginine starvation is induced by the upregulation of ASS enzymes (250). This current observation is in agreement with early reported findings on the repression of ASS and ASL reactions due to the presence of high levels of L-arginine in the culture media of cancer and mouse fibroblast cell lines (251).

Other metabolic tasks relevant for the physiology to be studied with the reduced model can be tested. However, it is important to correctly interpret the results of testing metabolic tasks in the context of the physiology represented by the reduced network (Figure B.4.6). Some tasks that should pass will fail simply because these parts of the network have not been kept within our core network during subsequent expansion as they were not required to ensure cell growth or because the pathways are lumped in a way that the route from the input compound to the compound produced becomes imperceptible.

After processing, the AllHuman data included in *RedHuman* consisted of 88 extracellular fluxes and 167 intracellular concentrations. From these, 36% of extracellular reactions were removed from *RedHuman*. In total, 20% of these extracellular fluxes and 16% of the metabolites with measured intracellular concentrations corresponded to metabolites that are not explicitly present in the *RedHuman* core network (see Table C.4.7). Metabolites that are not explicitly present but necessary for the cellular physiology under study are implicitly kept in the lumped

reactions that connect the core *RedHuman* network to each of the biomass building blocks (Table C.4.8). The remaining biomass building blocks without lumped reactions have their synthesis pathways explicitly present in the core network.

The AllHuman data used in the reduction of *RedHuman* and remained integrated into that model was replaced by the AllNormal and AllCancer data integrated separately to represent the normal and cancer physiologies, respectively. The maximum growth yield of each *RedHuman* representing normal and cancer physiologies is shown by the bar heights in Figure 4.8, where it is compared to the maximum growth rate measured for the cell lines included in the respective datasets (green horizontal line). The *RedHuman* for each physiology is able to reproduce the growth rate from the measurements. We note that since mammalian cells can uptake a multiplicity of substitute metabolites for which there may not be available measurements, the respective carbon pathways may not be included in the reduced model. This interchangeable use of substitutable carbon sources can lead to lower maximum growth yield. However, the flexibility of the reduction procedure combining redGEM and lumpGEM (205, 206) can be iteratively used to re-generate lumps by adding the required metabolites to the extracellular compartment and forcing their auxotrophy to generate the respective lumped pathways (Pipeline 2 in Figure B.4.8).

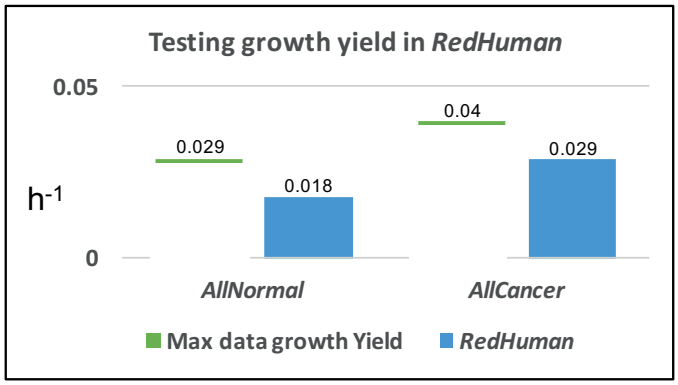


Figure 4.8 Maximum growth yield for each (blue bars) *RedHuman* with AllNormal and AllCancer data integrated representing normal and cancer physiologies, respectively, compared to the maximum growth rate measured for the cell lines included in the datasets (green horizontal line).

Both normal and cancer *RedHuman* models were analyzed with TFVA for a preliminary assessment of the differences in the feasible solution space of the model reactions given the two different physiologies. In Figure 4.9 we show the comparison of the flux variability bounds for some reactions between the two models to assess their potential for the analysis of two separate physiologies. Most of the observed variability is in the patterns of amino acid consumption (shown extracellular reactions). It is interesting to observe the blocking of certain reactions that take part in the urea cycle in the mitochondria for the cancer model, which could be related to an attempt in cancer to preserve L-arginine to be used in other biosynthetic reactions that support proliferation. We also observed that reactions involving the production of Nitric Oxide (NO) and its secretion are more flexible in the cancer physiology. This observation is also relevant in the context of cancer since NO levels have been associated with carcinogenesis and tumor growth progression (252).

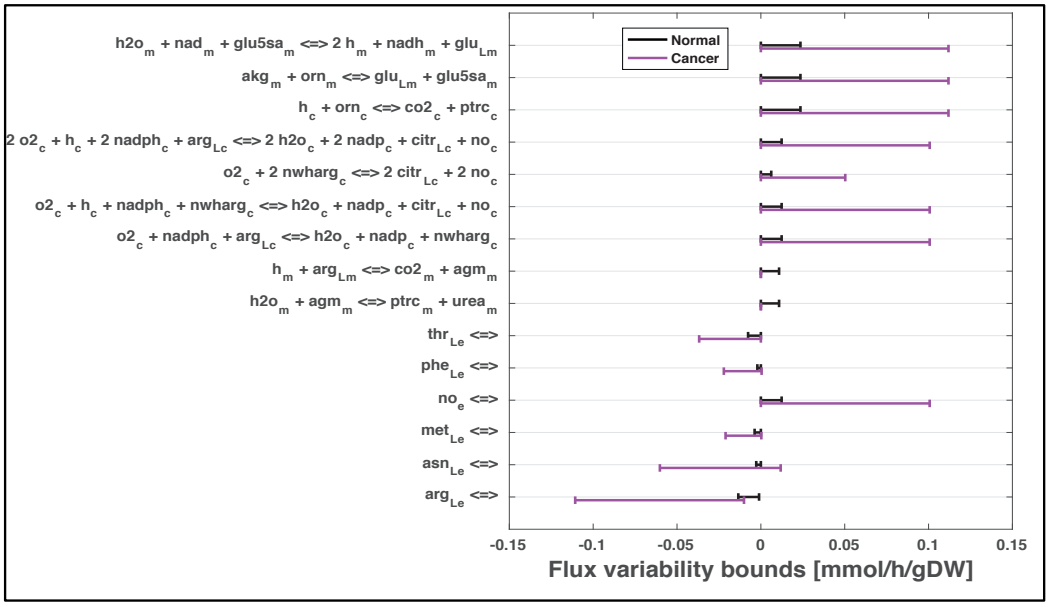


Figure 4.9 Comparison of feasible solution space of some reactions by performing Thermodynamics-based Flux Variability Analysis (TFVA) on *RedHuman* with AllCancer and AllNormal data.

4.4 Conclusion and path forward

In this chapter, we established the protocol for the reduction of a human metabolic model and the generation of tissue-specific models. We started with the human GEM Recon 2 v4 that has been processed and [re]-curated by the pipeline presented in Chapter 3 and demonstrated here its step-by-step reduction. Model reduction was performed using a semi-automatized workflow (redGEM and lumpGEM (205, 206)) that generates data-driven consistent models for normal and tissue-specific metabolic phenotypes, such that the models are specifically tailored to the physiology and conditions under study. *RedHuman*, the first thermodynamically feasible reduced model of human metabolism, was obtained by integrating into the network the fluxomics and metabolomics pertaining to normal human physiology and different tissue-specific cancers, as well as reaction thermodynamics constraints. The *RedHuman* generated had similar biosynthetic capabilities to the initial GEM, despite the size of the network being decreased by 3-fold in terms of reaction number which decreased the GEM associated complexity and network redundancies. Besides checking for the production of the biomass building blocks, the metabolic capabilities common to human metabolism, such as amino acid *de novo* synthesis, were also evaluated during validation stage. When the model was analyzed separately for the normal and cancer data, it showed the ability to differentiate between the two physiologies.

These reduced networks are the perfect scaffold to perform thermodynamics-based flux analysis with application in the study of key metabolic features associated with cancer metabolic reprogramming. They can be used to compare within and across tissue-specific cancers, as well as to perform comparative studies between the metabolic phenotype of a healthy cell type and cancer. In particular, since fluxomics for normal physiology is typically obtained for exponentially growing cells, such is the case of the datasets used in this work, this model can be applied to the comparative study between the metabolic reprogramming of growing normal cells with the one pertaining to highly-proliferating cancer cells. Transcriptomics data can also be added to the models to further differentiate the physiologies.

Furthermore, as discussed, the reduction pipeline implemented here is itself an analysis tool due to its intrinsic flexibility. It can be adapted to be performed in stages

where reaction lumping to the biomass building blocks are generated according to the tissue specific data that is integrated after generating the core network, thus allowing for direct comparison of intrinsic network heterogeneity among different tissues and conditions (Figure B.4.8). This allows for a systematic comparison between the metabolic states observed in normal growing cells and different cancers to identify the key metabolic features of the overflow metabolism (Warburg phenotype), which may be used as targets to induce reverse Warburg (as this ability has been observed in some cancers), as well as to identify cancer tissue specific targets.

As a final remark, we note that the pipeline described within Chapters 3 and 4, despite being presented for the human model, remains general and can be applied to the GEMs of any organism and for the study of other conditions/diseases.

Conclusion

In part I, we focused on the modeling and study of mRNA translation. The stochastic simulations of translation elongation in the context of an *E. coli* used a more detailed mathematical description of the ribosome kinetics during codon elongation, which was paramount to identify the relative importance of the determinants of elongation rate and unify the results from different computational and experimental studies on the subject.

We showed that the two factors that determine the speed of the ribosome along the mRNA strand are, by order of importance for *in vitro* conditions, the competition between the cognate and the non- and near-cognate tRNAs and the overall abundance of cognate tRNA interaction type (WC vs. WB). Interestingly, for *in vivo* conditions, tRNA competition becomes less important with respect to the cognate interaction type.

The simulations of heterologous translation of synonymous transcripts representing the same protein sequence, where codons were replaced by synonymous codons presenting the same amino acid and selected based on different criteria, showed that the transcript has a maximum elongation rate when its sequence is designed based on a derived equation that takes into account both determining factors and depends on the amount of free tRNA in the host cell. This work constitutes an important contribution to the field of synthetic biology as its results can be used to improve the design of sequences for heterologous protein production in pharmaceutical and biotech industries. It also strongly indicates that variability in tRNA abundance of different host organisms for protein production, such as CHO cell lines, can be a factor that influences the general productivities of these cell lines. The results from this study can thus be a motivator for the measurement of such quantities to screen ahead host cell lines to classify higher producers.

Part II focused on establishing a pipeline for generating consistent and thermodynamically feasible organism- and condition-specific reduced models from the GEMs of the respective organisms. Focusing on the human GEM Recon 2 v4, we developed tools and set up semi-automatized workflows (DRAMA) that allowed for a consistent, and more importantly, a reproducible processing of the model for –omics and compound thermodynamics parameters integration. This pipeline sets the standard for the curation of GEMs with total or partial missing annotation, a common problem of existing GEMs that inconveniences their ready use, with the purpose of providing the highest coverage on reaction thermodynamics for thermodynamics-based flux analysis, while it doubles as a guide for the processing of GEMs by non-experts in the field who wish to use them.

As demonstrated, this pipeline can be applied to the generation of tissue-specific cancer models for the study of cancer metabolic reprogramming, which can be used to study variability across different tissue-specific cancers, but also to explore differences pertaining to the Warburg effect by comparing cancer metabolism and healthy physiology under growth conditions.

This systematic pipeline that ranges from curation and processing of GEMs to building of reduced metabolic models opens the path to future studies where these reduced human metabolic networks are the perfect platform for more complex studies involving different cell types, similarly to what has been done for the gut microbioma (121, 122), where analysis with GEMs becomes cumbersome due to their size and data knowledge gaps. An interesting case study would be the analysis of the metabolic rewiring in the tumor microenvironment due to the presence of lactate shuttles between the cells in the tumor or in the surrounding tissue. Recent studies have shown that the exploration of these metabolic interactions within the tumor microenvironment could lead to the discovery of potential drug targets that would kill the more aggressive and resistant hypoxic tumor cells (253).

Appendix A Supplementary Texts and Methods

A.1 Supplementary Texts and Methods for Chapter 1

A.1.1 Modified ZH model equations

We extend the deterministic formulation of the ZH model of translation (50) into a modified ZH model that accounts for:

- Discrimination between near-cognate and non-cognate,
- Possibility for near-cognate (*nc*) misincorporation at proofreading stage,
- Proofreading kinetic step for cognate and near-cognate.

The schematic representation of the ribosome kinetic pathway is in Figure 1.4 of section 1.2.1 and the mass balance equations are the following:

$$\frac{dS_j^1}{dt} = V_{j,c}^{11} + V_{j,nc}^{11} + V_{j,c}^{r} + V_{j,nc}^{r} + V_{j,c}^{-1} + V_{j,nc}^{-1} + V_{j,non}^{-1} - V_{j,c}^{1} - V_{j,nc}^{1} - V_{j,non}^{1}$$

$$\frac{dS_{j,c}^2}{dt} = V_{j,c}^{1} + V_{j,c}^{-2} - V_{j,c}^{-1} - V_{j,c}^{2}$$

$$\frac{dS_{j,nc}^2}{dt} = V_{j,nc}^{1} + V_{j,nc}^{-2} - V_{j,nc}^{-1} - V_{j,nc}^{2}$$

$$\frac{dS_{j,non}^2}{dt} = V_{j,non}^{1} - V_{j,non}^{-1}$$

$$\frac{dS_{j,c}^3}{dt} = V_{j,c}^{2} + V_{j,c}^{-2} - V_{j,c}^{3}$$

$$\frac{dS_{j,nc}^3}{dt} = V_{j,nc}^{2} + V_{j,nc}^{-2} - V_{j,nc}^{3}$$

$$\frac{dS_{j,c}^4}{dt} = V_{j,c}^{3} - V_{j,c}^{4}$$

$$\frac{dS_{j,nc}^4}{dt} = V_{j,nc}^{3} - V_{j,nc}^{4}$$

$$\frac{dS_{j,c}^5}{dt} = V_{j,c}^{4} - V_{j,c}^{5} - V_{j,c}^{r}$$

$$\frac{dS_{j,nc}^5}{dt} = V_{j,nc}^{4} - V_{j,nc}^{5} - V_{j,nc}^{r}$$

$$\frac{dS_{j,c}^6}{dt} = V_{j,nc}^{5} + V_{j,nc}^{-6} - V_{j,nc}^{6}$$

$$\frac{dS_{j,nc}^6}{dt} = V_{j,nc}^{5} + V_{j,nc}^{-6} - V_{j,nc}^{6}$$

$$\frac{dS_{j,c}^9}{dt} = V_{j,c}^{6} - V_{j,c}^{9}$$

$$\frac{dS_{j,nc}^9}{dt} = k_{j,nc}^{6} - V_{j,nc}^{9}$$

$$\frac{dS_{j+1,c}^{11}}{dt} = V_{j+1,c}^{9} - V_{j+1,c}^{11}$$

$$\frac{dS_{j+1,nc}^{11}}{dt} = V_{j+1,nc}^{9} - V_{j+1,nc}^{11}$$

where S^s are the ribosome states (s) along the pathway, V_j^s are the reaction fluxes $\begin{cases} V_{j,bi}^1 = k_1^{bi} \cdot S_{j,bi}^1 . T_{bi}^f \\ V_{j,bi}^s = k_s^{bi} \cdot S_{j,bi}^s, \quad s = [2, \dots, 6, 9, 11] \end{cases}$, bi is the binding interaction (cognate (c) or near-

cognate (*nc*)), *s* represents the state number [1,...,6,9,11], and *j* is the codon number that is being decoded at the ribosome A-site. Since less intermediate steps were used in ZH model, we introduced a discontinuity in the index of the ribosome states in our modified ZH model to establish a connection with further model extensions (Table C.1.1).

The system of mass balance equations above can be simplified by writing only one equation in terms of a state that is a combination of all the intermediate states. We choose state 9 (the ribosome translocation state) as our new lumped state. The new system will be given by

$$\frac{dS_j^{lumped}}{dt} = V_{in} - V_{out} = V_{j,c}^9 + V_{j,nc}^9 - k_{eff}^j \cdot S_j^{lumped}, \tag{A.1.1}$$

where

$$S_j^{lumped} = \sum_s S_j^s. \tag{A.1.2}$$

For a system at steady state, the effective codon elongation rate constant can be written as

$$k_{eff}^j = \frac{V_{j,c}^9 + V_{j,nc}^9}{\sum_s S_j^s}. \tag{A.1.3}$$

Solving the mass balance equations at steady state all states can be re-written in function of state 9 and k_{eff}^j can thus be expressed in the final form

$$k_{eff}^j = \frac{1 + k_5^{nc}\alpha_5^{nc} T_{nc,j}^f \alpha_{init-selec}^j \left(1 + \frac{k_{rej}^c}{k_5^c}\right)}{\alpha_{comp}^j + \alpha_c + \alpha_{init-selec}^j \left(1 + \frac{k_{rej}^c}{k_5^c}\right)\alpha_{mis-inc}^j}, \tag{A.1.4}$$

where $T_{bi,j}^f$ are the tRNA concentrations for different binding interaction types at codon j.

The terms in the expression are defined below:

<u>Cognate WC term</u>:

$$\alpha_c^j = \left(1 + \frac{k_{rej}^c}{k_5^c}\right)\left(\alpha_{init-selec}^j + \alpha_2^c + \alpha_3^c + \alpha_4^c\right) + \left(\alpha_5^c + \alpha_{6,f}^c + \alpha_{6,b}^c + \alpha_9^c + \alpha_{11}^c\right) \tag{A.1.5}$$

Initial selection term:

$$\alpha_{init-selec}^{j} = \frac{1}{k_1^c T_{c,j}^f}\left[\frac{(k_{-1}^c + k_2^c)(k_3^c + k_{-2}^c)}{k_3^c k_2^c} - \frac{k_{-2}^c}{k_3^c}\right] \quad \text{(A.1.6)}$$

Competition term:

$$\alpha_{comp}^{j} = \alpha_{init-selec}^{j}\left(1 + \frac{k_{rej}^c}{k_5^c}\right)\left(\begin{array}{c}\alpha_2^{non} T_{non,j}^f + \\ (\alpha_2^{nc} + \alpha_3^{nc} + \alpha_4^{nc} + \alpha_5^{nc}) T_{nc,j}^f\end{array}\right) \quad \text{(A.1.7)}$$

Mis-incorporation term (near-cognate proofreading):

$$\alpha_{mis-inc}^{j} = k_5^{nc}\alpha_5^{nc}\left(\alpha_{6,f}^{nc} + \alpha_{6,b}^{nc} + \alpha_9^{nc} + \alpha_{11}^{nc}\right) T_{nc,j}^f \quad \text{(A.1.8)}$$

Other terms:

Non-cognate

$$\alpha_2^{non} = \frac{k_1^{non}}{k_{-1}^{non}} \quad \text{(A.1.9)}$$

Cognate WC

$$\alpha_2^c = \frac{\left(k_{-2}^c + k_3^c\right)}{k_2^c k_3^c}$$

$$\alpha_3^c = \frac{1}{k_3^c}$$

$$\alpha_4^c = \frac{1}{k_4^c}$$

$$\alpha_5^c = \frac{1}{k_5^c}$$

$$\alpha_{6,f}^c = \frac{1}{G^{(f)} \cdot k_6^c}$$

$$\alpha_{6,b}^c = \frac{k_{-6}^c}{\left(G^{(f)} \cdot k_9^c \cdot k_6^c \cdot W_{j+1}\right)}$$

$$\alpha_9^c = \frac{1}{k_9^c \cdot W_{j+1}}$$

$$\alpha_{11}^c = \frac{1}{k_{11}^c}$$

Near-cognate

$$\alpha_2^{nc} = \frac{k_1^{nc}\left(k_3^{nc} + k_{-2}^{nc}\right)}{k_3^{nc}k_{-1}^{nc} + k_{-1}^{nc}k_{-2}^{nc} + k_3^{nc}k_2^{nc}}$$

$$\alpha_3^{nc} = \frac{k_1^{nc}k_2^{nc}}{k_3^{nc}k_{-1}^{nc} + k_{-1}^{nc}k_{-2}^{nc} + k_3^{nc}k_2^{nc}}$$

$$\alpha_4^{nc} = \frac{k_3^{nc}k_1^{nc}k_2^{nc}}{k_4^{nc}\left(k_3^{nc}k_{-1}^{nc} + k_{-1}^{nc}k_{-2}^{nc} + k_3^{nc}k_2^{nc}\right)}$$

$$\alpha_5^{nc} = \frac{k_3^{nc}k_1^{nc}k_2^{nc}}{\left(k_5^{nc} + k_{rej}^{nc}\right)\left(k_3^{nc}k_{-1}^{nc} + k_{-1}^{nc}k_{-2}^{nc} + k_3^{nc}k_2^{nc}\right)}$$

$$\alpha_{6,f}^{nc} = \frac{1}{k_6^{nc} \cdot G^{(f)}}$$

$$\alpha_{6,b}^{nc} = \frac{k_{-6}^{nc}}{G^{(f)} \cdot k_6^{nc} \cdot k_9^{nc} \cdot W_{j+1}}$$

$$\alpha_9^{nc} = \frac{1}{k_9^{nc} \cdot W_{j+1}}$$

$$\alpha_{11}^{nc} = \frac{1}{k_{11}^{nc}}$$

The term W_{j+1} present in equations $\alpha_{6,b}$ and α_9 represents the conditional probability that codon $j+1$ is free given that the ribosome with A site in codon j is ready for translocation. This term was defined initially in (39) and later on used in the ZH model, and is given by

$$W_{j+1} = \begin{cases} \frac{1-\sum_{k=1}^{L_R} x_{j+k}}{1-\sum_{k=1}^{L_R-1} x_{j+k}}, & if\ 1 \le j \le n - L_R \\ 1, & otherwise \end{cases} \qquad \text{(A.1.10)}$$

where L_R is the length of the ribosome in terms of number of codons it occupies on the mRNA, n is the total number of codons in the mRNA transcript, and x_{j+k} is the fractional codon occupancy of codon $j+k$ by a ribosome A site.

A.1.2 Reproduction of experiments with modified ZH model

The modified ZH model is used for reproducing the experimental results in (49). The dynamic simulations were performed using the ribosome kinetic rates summarized in (53) for the translation of a mRNA segment of the form auguuuuuu(...)uaa, with 26 UUU codons in between the start and stop codons. Experiments were performed with 0.2 μM of isolated cognate or near cognate TC so that no competition takes place and 2.8 μM of initiation complexes (ICs). Since the model is built for a constant supply of TC and the experiments use instead an excess of $IC/mRNA$ with respect to the TCs, we replaced T_{bi}^{f} in equation $V_{n,bi}^{1} = k_1^{bi} \cdot S_{n,bi}^{1} \cdot T_{bi}^{f}$ by $M = [IC] = 2.8$ μM and, instead of referring to the ribosomal states S^i, we refer to the TC states in the ribosome ($V_{n,bi}^{1} = k_1^{bi} \cdot S_{n,bi}^{1} \cdot M$). We simulated the dynamic model during 20 and 50 seconds for cognate and near cognate, respectively. The simulation was initialized at the translocation step of the start codon, which is state one of the next codon, in order to reproduce the experimental conditions in which the start codon has been recognized and it is sitting on the ribosome P-site. Proofread aa-tRNAs do not add to the bulk of available TCs because in the experiments there is no elongation factor EF-Ts in the medium to promote the assembly of the tRNAs with the EF-Tu and GTP to form a TC. The curves in Figure 1.5 and Figure 1.6 were obtained by integrating the flux resulting from the kinetic step corresponding to the GTP hydrolysis $\left(\frac{dS_{n=2}^{3}}{dt} = k_3 S_{n=2}^{3}\right)$ or the accommodation $\left(\frac{dS_{n=2}^{5}}{dt} = k_5 S_{n=2}^{5}\right)$ during translation of the second codon, respectively.

A.1.3 Sobol's Method for GSA

Global sensitivity analysis (GSA) methods quantify the influence of the uncertainty of the model parameters on the variance of the model output. GSA methods have the advantage of dealing with models regardless of any assumption in linearity. Furthermore, they allow for an exploration of the entire parameter space and take into account interactions among the parameters.

The Sobol's method is a variance-based approach that uses a Monte-Carlo based numerical procedure to sample the parameter space with a quasi-random number

generator., which estimates the individual parameter effects on the model output variance (a detailed description can be found in (254)).

Briefly, for a model output described as a function of its parameters (X_i), $Y = f(X_1, \dots, X_k)$, the total variance of the model output

$$V(Y) = \sum_i V_i + \sum_i \sum_{j>i} V_{ij} + \cdots + V_{12\dots k}, \quad \text{(A.1.11)}$$

can be decomposed into the sum of the first order conditional variance of each parameter i (V_i) and the second order conditional variance of each parameter combination ($V_{i,j}, i \neq j$). Parameter sensitivity is quantified by two derived measures: the main effect and the total effect.

The first order global sensitivity index (or **main effect**) is defined as

$$S_i = \frac{V_i}{V(Y)}. \quad \text{(A.1.12)}$$

This index gives a measure of how much the change in one parameter influences the model output.

The **total effec**t index

$$S_{T_i} = S_i + \sum_{j \neq i} S_{ij} + \cdots + S_{12\dots k}, \quad \text{(A.1.13)}$$

where S_{ijk} represents the joint effect contribution of parameters $X_{i,j,k}$ to the output variance and hence accounts for the **total contribution** of each parameter to the model output.

A.2 Supplementary Texts and Methods for Chapter 2

A.2.1 Cell composition in ribosome and tRNA molecules

We estimated the total number of tRNA molecules per cell ($tRNA^T$) for the growth rates 0.7, 1.07, and 1.6 h^{-1} from the fit of the total number of tRNA molecules per cell in function of other growth rates (0.6, 1.0, 1.5, 2, 2.5 h^{-1}) that were reported in (58) for an exponentially growing *E. coli* cell. For the growth rate 0.4 h^{-1} this fitting step was not necessary since the number of molecules for all the tRNA species was directly reported

in (59) at this growth rate, along with their concentrations. The same fitting procedure and source of data mentioned above was used to estimate the total number of ribosomes per cell (R^T) for all the four growth rates of interest (0.4, 0.7, 1.07, and 1.6 h^{-1}). Fittings can be found in Figure B.2.1. The concentrations for all tRNA species in *E. coli* at these growth rates were obtained from the experiments reported in (59). The tRNA isoacceptors Gly1-Gly2 and Ile1-Ile2 were treated collectively in (59) and we proceeded to split their values according to the ratio of their gene copy number, which are Gly1:Gly2=1:1 and Ile1:Ile2=3:1 (82). The cell volume (V_{cell}) was then computed for each growth rate using

$$V_{cell} = \frac{tRNA^T}{N_A \cdot [tRNA^T]}, \quad \text{(A.2.1)}$$

where $[tRNA^T]$ is the total concentration of tRNA molecules at the given growth rate and N_A is the Avogadro constant. With the estimated cell volumes and the concentrations of each tRNA species, we calculated the number of molecules for each tRNA species at each growth rate. The $tRNA^T$ and R^T values obtained from the fittings and the computed V_{cell} at each growth rate are summarized in Table C.2.5.

A.2.2 mRNA sequences present in the simulated cell

Similarly to what was done for $tRNA^T$ and R^T, we estimated the mRNA synthesis rate per cell for the growth rates (0.7, 1.07, and 1.6 h^{-1}) from the fitting of the mRNA synthesis rate in function of growth rate reported in (58) for an exponentially growing *E. coli* cell (see Figure B.2.2 for fitting and Table C.2.5 for the values). The average number of mRNA copies per *E. coli* (M_T) was computed for each growth rate of interest (0.4, 0.7, 1.07, and 1.6 h^{-1}) with the following expression

$$M_T = \frac{\upsilon_{mRNA} \cdot \tau_{mRNA}}{nt_{mRNA}}, \quad \text{(A.2.2)}$$

where υ_{mRNA} is the rate of mRNA synthesis per cell, τ_{mRNA} is the average functional life of mRNA, and nt_{mRNA} is the average number of nucleotides for the mRNA sequences in *E. coli*. The average functional life of mRNA is assumed both in (58) and here to be 1 min and independent of growth rate, as estimated from pulse-labeling experiments in (255). A value of 317 codons was obtained for the average mRNA length for *E. coli* ($L_{database}$)

with a standard deviation of 213 codons. This computation was based on averaging the length of the protein coding regions from all the mRNA species in EcoGene 3.0 database for the strain *E. coli K12* (81).

Since we lack data on the mRNA sequences and respective copy numbers expressed at each of the growth rates under study, we constructed the mRNA pools of the cell at each condition by formulating a homogeneity criterion based on the fact that *E. coli* expresses mRNA in low copy number (78). This criterion assumes that the mRNA pools are qualitatively similar across the four growth rates and enforces them to approximate both the average mRNA length and the codon usage frequency (CU) of *E. coli.* In Table S5 there is a comparison between the mRNA expression in *E. coli* at low (79) and high (80) growth rates. The statistics over the mRNA copy number distributions at both conditions show that most mRNAs are expressed in very small amounts with not so frequent occurrences of bursts in mRNA expression levels. Furthermore, the average codon length of the mRNA sequences expressed at each condition is similar to the average mRNA length representing the whole *E. coli* genome and the average of the relative deviations between mCU and CU for each codon species is 15% and 23%, respectively, for low and high growth conditions. We obtained nt_{mRNA} by multiplying $\bar{L}_{database}$ by 3 (number of nucleotides in each codon), which was then assumed to be constant across growth rates (see Table C.2.5 for values).

In order to obtain a homogenous mRNA pool for each growth rate, we started by selecting a subset of mRNA species from EcoGene 3.0 database. In this subset, 52% of the sequences were classified as essential genes. The total number of mRNA species present in this subset (289 species) was chosen to match the M_T at growth rate 0.4h^{-1} such that it contains exactly 1 copy for each mRNA species at this growth rate. The selection criteria for the choice of this subset of mRNA species was based on

1. Reaching a relative deviation smaller than 0.5% between the average mRNA length of the subset and $\bar{L}_{database}$,
2. And at the same time enforcing the relative deviation between CU and mCU for each codon to present an average and standard deviation across all codons both smaller than 10%.

We chose to control both the average and the standard deviation for the second criterion so that we obtained a more homogeneous set of deviations between CU and mCU for each codon in the attempt to keep the mCU values close to *E. coli* CU.

In order to set the mRNA pool for the remaining three growth rates, we performed an iterative process for each growth rate focusing on the two selection criteria presented above. The number and type of mRNA species chosen for the growth rate $0.4h^{-1}$ was maintained for the remaining three growth rates, but their copy numbers were increased to match the target number of mRNA molecules for each of these growth rates as presented in Table C.2.5. The pools were at first increased homogeneously, i.e., each mRNA species was increased by the same amount of copies until the total number became as close as possible to the target M_T. Subsequently, a group of different mRNA species that matched the number of mRNA copies missing to reach the target M_T was randomly selected to have its copy number increased. This random selection was repeated for each growth rate until criteria 1 and 2 from above were fulfilled, allowing us to construct mRNA pools that are qualitatively similar across the four growth rates and with a mCU that approximates the *E. coli* CU (Figure B.2.3).

A.2.3 Derivation of full deterministic equation for computation of codon elongation rate

We further extend the deterministic formulation of the modified ZH model of translation (section A.1.1):

- Discrimination between cognate Watson-Crick (*WC*) and cognate wobble (*WB*),
- Possibility for near-cognate (*nc*) misincorporation at proofreading stage,
- Proofreading kinetic step for cognates (*WC* and *WB*) and near-cognate.

These deterministic equations are developed for a low ribosomal density, where the effects of the ribosome queuing can be ignored. The schematic representation of the ribosome kinetic pathway is the one in Figure 2.1 of section 2.2.1.1 and the mass balance equations are the following:

$$
\begin{aligned}
\frac{dS_j^1}{dt} &= V_{j,WC}^{11} + V_{j,WB}^{11} + V_{j,nc}^{11} + V_{j,WC}^{r} + V_{j,WB}^{r} + \\
&+ V_{j,nc}^{r} + V_{j,WC}^{-1} + V_{j,WB}^{-1} + V_{j,nc}^{-1} + V_{j,non}^{-1} - \\
&- V_{j,WC}^{1} - V_{j,WB}^{1} - V_{j,nc}^{1} - V_{j,non}^{1} \\
\frac{dS_{j,WC}^2}{dt} &= V_{j,WC}^{1} + V_{j,WC}^{-2} - V_{j,WC}^{-1} - V_{j,WC}^{2} \\
\frac{dS_{j,WB}^2}{dt} &= V_{j,WB}^{1} + V_{j,WB}^{-2} - V_{j,WB}^{-1} - V_{j,WB}^{2} \\
\frac{dS_{j,nc}^2}{dt} &= V_{j,nc}^{1} + V_{j,nc}^{-2} - V_{j,nc}^{-1} - V_{j,nc}^{2} \\
\frac{dS_{j,non}^2}{dt} &= V_{j,non}^{1} - V_{j,non}^{-1} \\
\frac{dS_{j,WC}^3}{dt} &= V_{j,WC}^{2} + V_{j,WC}^{-2} - V_{j,WC}^{3} \\
\frac{dS_{j,WB}^3}{dt} &= V_{j,WB}^{2} + V_{j,WB}^{-2} - V_{j,WB}^{3} \\
\frac{dS_{j,nc}^3}{dt} &= V_{j,nc}^{2} + V_{j,nc}^{-2} - V_{j,nc}^{3} \\
\frac{dS_{j,WC}^4}{dt} &= V_{j,WC}^{3} - V_{j,WC}^{4} \\
\frac{dS_{j,WB}^4}{dt} &= V_{j,WB}^{3} - V_{j,WB}^{4} \\
\frac{dS_{j,nc}^4}{dt} &= V_{j,nc}^{3} - V_{j,nc}^{4} \\
\frac{dS_{j,WC}^5}{dt} &= V_{j,WC}^{4} - V_{j,WC}^{5} - V_{j,WC}^{r} \\
\frac{dS_{j,WB}^5}{dt} &= V_{j,WB}^{4} - V_{j,WB}^{5} - V_{j,WB}^{r} \\
\frac{dS_{j,nc}^5}{dt} &= V_{j,nc}^{4} - V_{j,nc}^{5} - V_{j,nc}^{r} \\
\frac{dS_{j,WC}^6}{dt} &= V_{j,WC}^{5} + V_{j,WC}^{-6} - V_{j,WC}^{6} \\
\frac{dS_{j,WB}^6}{dt} &= V_{j,WB}^{5} + V_{j,WB}^{-6} - V_{j,WB}^{6} \\
\frac{dS_{j,nc}^6}{dt} &= V_{j,nc}^{5} + V_{j,nc}^{-6} - V_{j,nc}^{6} \\
\frac{dS_{j,WC}^7}{dt} &= V_{j,WC}^{6} - V_{j,WC}^{-6} - V_{j,WC}^{7} \\
\frac{dS_{j,WB}^7}{dt} &= V_{j,WB}^{6} - V_{j,WB}^{-6} - V_{j,WB}^{7} \\
\frac{dS_{j,nc}^7}{dt} &= V_{j,nc}^{6} - V_{j,nc}^{-6} - V_{j,nc}^{7}
\end{aligned}
\tag{A.2.3}
$$

$$
\begin{aligned}
\frac{dS^{8}_{j,WC}}{dt} &= V^{7}_{j,WC} - V^{8}_{j,WC} \\
\frac{dS^{8}_{j,WB}}{dt} &= V^{7}_{j,WB} - V^{8}_{j,WB} \\
\frac{dS^{8}_{j,nc}}{dt} &= V^{7}_{j,nc} - V^{8}_{j,nc} \\
\frac{dS^{9}_{j,WC}}{dt} &= V^{8}_{j,WC} - V^{9}_{j,WC} \\
\frac{dS^{9}_{j,WB}}{dt} &= V^{8}_{j,WB} - V^{9}_{j,WB} \\
\frac{dS^{9}_{j,nc}}{dt} &= V^{8}_{j,nc} - V^{9}_{j,nc} \\
\frac{dS^{10}_{j+1,WC}}{dt} &= V^{9}_{j,WC} - V^{10}_{j+1,WC} \\
\frac{dS^{10}_{j+1,WB}}{dt} &= V^{9}_{j,WB} - V^{10}_{j+1,WB} \\
\frac{dS^{10}_{j+1,nc}}{dt} &= V^{9}_{j,nc} - V^{10}_{j+1,nc} \\
\frac{dS^{11}_{j+1,WC}}{dt} &= V^{10}_{j+1,WC} - V^{11}_{j+1,WC} \\
\frac{dS^{11}_{j+1,WB}}{dt} &= V^{10}_{j+1,WB} - V^{11}_{j+1,WB} \\
\frac{dS^{11}_{j+1,nc}}{dt} &= V^{10}_{j+1,nc} - V^{11}_{j+1,nc}
\end{aligned}
\qquad \text{(A.2.3 cont.)}
$$

where S^s are the ribosome states (s) along the pathway, V_n^s are the reaction fluxes $\begin{pmatrix} V^1_{j,bi}=k^{bi}_1 \cdot S^1_{j,bi} \cdot T^f_{bi} \\ V^s_{j,bi}=k^{bi}_s \cdot S^s_{j,bi}, \quad s=2,3,\ldots,11 \end{pmatrix}$, bi is the binding interaction (cognate WC (WC), cognate WB (WB), near-cognate (nc), or non-cognate (non)), s represents the state number from 1 to 11, and j is the codon number that is being decoded at the ribosome A-site. In state 9, ribosome translocation takes place and a new codon is placed in the A-site.

This system of mass balance equations can be simplified by writing only one equation in terms of a state that is a combination of all the intermediate states. We choose state 9 (the ribosome translocation state) as our new lumped state. The new system will be given by

$$\frac{dS_j^{lumped}}{dt} = V_{in} - V_{out} = V_{j,WC}^9 + V_{j,WB}^9 + V_{j,nc}^9 - k_{eff}^j \cdot S_j^{lumped}, \quad \text{(A.2.4)}$$

where

$$S_j^{lumped} = \sum_s S_j^s. \quad \text{(A.2.5)}$$

For a system at steady state, the effective codon elongation rate constant can be written as

$$k_{eff}^j = \frac{V_{j,WC}^9 + V_{j,WB}^9 + V_{j,nc}^9}{\sum_s S_j^s}. \quad \text{(A.2.6)}$$

Solving the mass balance equations at steady state all states can be re-written in function of state 9 and k_{eff}^j can thus be expressed in the final form

$$k_{eff}^j = \frac{T_{WC,j}^f + k_5^{WB}\alpha_5^{WB}T_{WB,j}^f\alpha_{init-selec}\left(1+\frac{k_{rej}^{WC}}{k_5^{WC}}\right) + k_5^{nc}\alpha_5^{nc}T_{nc,j}^f\alpha_{init-selec}\left(1+\frac{k_{rej}^{WC}}{k_5^{WC}}\right)}{\alpha_{comp}^j + \alpha_{WC}^j + \alpha_{init-selec}\left(1+\frac{k_{rej}^{WC}}{k_5^{WC}}\right)\left(\alpha_{WB-inc}^j + \alpha_{mis-inc}^j\right)}, \quad \text{(A.2.7)}$$

where $T_{bi,j}^f$ are the tRNA concentrations for different binding interaction types at codon j.

The terms in the expression are defined below.

Cognate WC term:

$$\alpha_{WC}^j = \left(1 + \frac{k_{rej}^{WC}}{k_5^{WC}}\right)\left(\alpha_{init-selec} + \alpha_2^{WC}T_{WC,j}^f + \alpha_3^{WC}T_{WC,j}^f + \alpha_4^{WC}T_{WC,j}^f\right) +$$

$$\left(\alpha_5^{WC} + \alpha_{6,f}^{WC} + \alpha_{6,b}^{WC} + \alpha_7^{WC} + \alpha_8^{WC} + \alpha_9^{WC} + \alpha_{10}^{WC} + \alpha_{11}^{WC}\right)T_{WC,j}^f \quad \text{(A.2.8)}$$

Initial selection term:

$$\alpha_{init-selec} = \frac{1}{k_1^{WC}}\left[\frac{\left(k_{-1}^{WC} + k_2^{WC}\right)\left(k_3^{WC} + k_{-2}^{WC}\right)}{k_3^{WC}k_2^{WC}} - \frac{k_{-2}^{WC}}{k_3^{WC}}\right] \quad \text{(A.2.9)}$$

Competition term:

$$\alpha_{comp}^{j} = \alpha_{init-selec}\left(1 + \frac{k_{rej}^{WC}}{k_{5}^{WC}}\right) \cdot \begin{bmatrix} \alpha_{2}^{non} T_{non,j}^{f} + (\alpha_{2}^{nc} + \alpha_{3}^{nc} + \alpha_{4}^{nc} + \alpha_{5}^{nc}) T_{nc,j}^{f} + \\ (\alpha_{2}^{WB} + \alpha_{3}^{WB} + \alpha_{4}^{WB} + \alpha_{5}^{WB}) T_{WB,j}^{f} \end{bmatrix} \quad \text{(A.2.10)}$$

WB-incorporation term (cognate WB proofreading):

$$\alpha_{WB-inc}^{j} = k_{5}^{WB} \alpha_{5}^{WB} \begin{pmatrix} \alpha_{6,f}^{WB} + \alpha_{6,b}^{WB} + \alpha_{7}^{WB} + \\ \alpha_{8}^{WB} + \alpha_{9}^{WB} + \alpha_{10}^{WB} + \alpha_{11}^{WB} \end{pmatrix} T_{WB,j}^{f} \quad \text{(A.2.11)}$$

Mis-incorporation term (near-cognate proofreading):

$$\alpha_{mis-inc}^{j} = k_{5}^{nc} \alpha_{5}^{nc} \begin{pmatrix} \alpha_{6,f}^{nc} + \alpha_{6,b}^{nc} + \alpha_{7}^{nc} + \\ \alpha_{8}^{nc} + \alpha_{9}^{nc} + \alpha_{10}^{nc} + \alpha_{11}^{nc} \end{pmatrix} T_{nc,j}^{f} \quad \text{(A.2.12)}$$

Other terms:

Non-cognate

$$\alpha_{2}^{non} = \frac{k_{1}^{non}}{k_{-1}^{non}} \quad \text{(A.2.13)}$$

Cognate WC

$$\alpha_2^{WC} = \frac{\left(k_{-2}^{WC} + k_3^{WC}\right)}{k_2^{WC} k_3^{WC}}$$

$$\alpha_3^{WC} = \frac{1}{k_3^{WC}}$$

$$\alpha_4^{WC} = \frac{1}{k_4^{WC}}$$

$$\alpha_5^{WC} = \frac{1}{k_5^{WC}}$$

$$\alpha_{6,f}^{WC} = \frac{1}{G^{(f)} \cdot k_6^{WC}}$$

$$\alpha_{6,f}^{WC} = \frac{k_{-6}^{WC}}{\left(G^{(f)} \cdot k_7^{WC} \cdot k_6^{WC}\right)}$$

$$\alpha_7^{WC} = \frac{1}{k_7^{WC}}$$

$$\alpha_8^{WC} = \frac{1}{k_8^{WC}}$$

$$\alpha_9^{WC} = \frac{1}{k_9^{WC}}$$

$$\alpha_{10}^{WC} = \frac{1}{k_{10}^{WC}}$$

$$\alpha_{11}^{WC} = \frac{1}{k_{11}^{WC}}$$

Cognate WB

$$\alpha_2^{WB} = \frac{k_1^{WB}\left(k_3^{WB} + k_{-2}^{WB}\right)}{k_3^{WB} k_{-1}^{WB} + k_{-1}^{WB} k_{-2}^{WB} + k_3^{WB} k_2^{WB}}$$

$$\alpha_3^{WB} = \frac{k_1^{WB} k_2^{WB}}{k_3^{WB} k_{-1}^{WB} + k_{-1}^{WB} k_{-2}^{WB} + k_3^{WB} k_2^{WB}}$$

$$\alpha_4^{WB} = \frac{k_3^{WB} k_1^{WB} k_2^{WB}}{k_4^{WB}\left(k_3^{WB} k_{-1}^{WB} + k_{-1}^{WB} k_{-2}^{WB} + k_3^{WB} k_2^{WB}\right)}$$

$$\alpha_5^{WB} = \frac{k_3^{WB} k_1^{WB} k_2^{WB}}{\left(k_5^{WB} + k_{rej}^{WB}\right)\left(k_3^{WB} k_{-1}^{WB} + k_{-1}^{WB} k_{-2}^{WB} + k_3^{WB} k_2^{WB}\right)}$$

$$\alpha_{6,f}^{WB} = \frac{1}{k_6^{WB} \cdot G^{(f)}}$$

$$\alpha_{6,b}^{WB} = \frac{k_{-6}^{WB}}{k_6^{WB} \cdot k_7^{WB} \cdot G^{(f)}}$$

$$\alpha_7^{WB} = \frac{1}{k_7^{WB}}$$

$$\alpha_8^{WB} = \frac{1}{k_8^{WB}}$$

$$\alpha_9^{WB} = \frac{1}{k_9^{WB}}$$

$$\alpha_{10}^{WB} = \frac{1}{k_{10}^{WB}}$$

$$\alpha_{11}^{WB} = \frac{1}{k_{11}^{WB}}$$

Near-cognate

$$\alpha_2^{nc} = \frac{k_1^{nc}\left(k_3^{nc} + k_{-2}^{nc}\right)}{k_3^{nc} k_{-1}^{nc} + k_{-1}^{nc} k_{-2}^{nc} + k_3^{nc} k_2^{nc}}$$

$$\alpha_3^{nc} = \frac{k_1^{nc} k_2^{nc}}{k_3^{nc} k_{-1}^{nc} + k_{-1}^{nc} k_{-2}^{nc} + k_3^{nc} k_2^{nc}}$$

$$\alpha_4^{nc} = \frac{k_3^{nc} k_1^{nc} k_2^{nc}}{k_4^{nc}\left(k_3^{nc} k_{-1}^{nc} + k_{-1}^{nc} k_{-2}^{nc} + k_3^{nc} k_2^{nc}\right)}$$

$$\alpha_5^{nc} = \frac{k_3^{nc} k_1^{nc} k_2^{nc}}{\left(k_5^{nc} + k_{rej}^{nc}\right)\left(k_3^{nc} k_{-1}^{nc} + k_{-1}^{nc} k_{-2}^{nc} + k_3^{nc} k_2^{nc}\right)}$$

$$\alpha_{6,f}^{nc} = \frac{1}{k_6^{nc} \cdot G^{(f)}}$$

$$\alpha_{6,b}^{nc} = \frac{k_{-6}^{nc}}{k_6^{nc} \cdot k_7^{nc} \cdot G^{(f)}}$$

$$\alpha_7^{nc} = \frac{1}{k_7^{nc}}$$

$$\alpha_8^{nc} = \frac{1}{k_8^{nc}}$$

$$\alpha_9^{nc} = \frac{1}{k_9^{nc}}$$

$$\alpha_{10}^{nc} = \frac{1}{k_{10}^{nc}}$$

$$\alpha_{11}^{nc} = \frac{1}{k_{11}^{nc}}$$

Inserting the values of the kinetic rate constants from Table C.2.1 we obtained an expression to compute k_{eff}^{j} for each codon *j*

$$k_{eff}^{j} = \frac{T_{WC,j}^{f} + 0.5884 \cdot T_{WB,j}^{f} + 2.6233 \cdot 10^{-4} \cdot T_{nc,j}^{f}}{0.0104[\mu M \cdot s] + 0.4556[s] \cdot T_{WC,j}^{f} + 0.0613[s] \cdot T_{nc,j}^{f} + 0.0171[s] \cdot T_{non,j}^{f}}\ [s^{-1}], \qquad \text{(A.2.14)}$$

where the variables are the free *WC*, *WB*, near-cognate and non-cognate tRNA concentrations to codon *j*.

A.3 Supplementary Texts and Methods for Chapter 3

A.3.1 Database web services protocols & system requirements

Operating system used was MacOSX, but Windows and Unix would have similar requirements and implementation.

Localhost simulation of BiGG database in MacOSX requires installation of Docker (https://docs.docker.com/engine/getstarted/step_one/#step-1-get-docker) and bigg_docker (https://github.com/psalvy/bigg-docker) (see links for installation).

For access to local server databases in SQL system requires from within MATLAB the installation of mysql-connector-java-5.1.39. For functions running parallel queries on SQL databases, system requires installation of mysql (https://dev.mysql.com/doc/refman/5.6/en/osx-installation-pkg.html) and Python 2.7 (https://www.python.org/download/releases/2.7/).

Web services running on SOAP API require installation of SOAP:Lite (http://search.cpan.org/dist/SOAP-Lite/). Instructions for installation on MacOSX can be found here (http://www.soaplite.com/2003/06/installation_in.html). Perl (perl-5.22.0) was used for writing the scripts for access and queries in ChEBI, which can be called from within MATLAB.

Web services running on REST API do not require any specific installations. Easily programed in MATLAB with either *urlread.m* or *webread.m* functions.

A.3.2 Databases in local server

ModelSEED database information was downloaded from https://github.com/ModelSEED/. ModelSEEDDatabase on March 17th 2017. Txt files containing the aliases between ModelSEED compounds and other databases were parsed in MATLAB to a seeming database structure and uploaded to our local server in

SQL. Metabolite structure was reused from previous ModelSEED database in our server and the new entries updated by converting InChI into molfiles and SMILE. Molconvert was used for InChI conversion into MOL and SMILE, Marvin 16.7.4, 2016, ChemAxon (http://www.chemaxon.com) (161).

HMDB database information was download December, 28th 2016 (last release on website at the time of writing dates of 2017-05-14). SDF file for metabolite structures (MOL,InChI, SMILES) and XML file for all metabolite information with compound synonym names and cross references were both parsed in MATLAB to a seeming database structure and uploaded to our local server in SQL.

A.4 Supplementary Texts and Methods for Chapter 4

A.4.1 Studies with GEMs

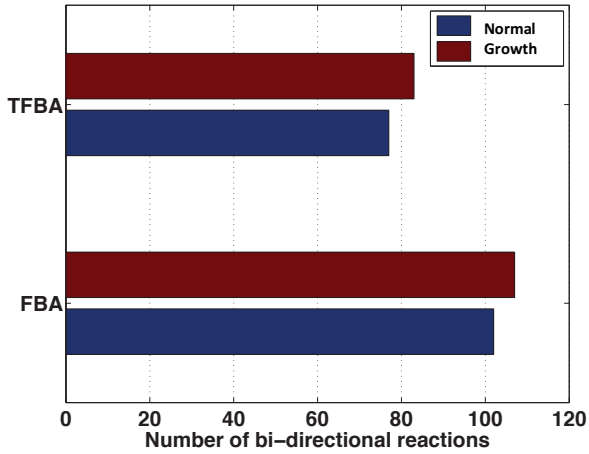
We derived a reduced mammalian metabolic network from Recon 1 using redGEM (205). The network focused on central carbon metabolic (10). In total the reduced model comprised 278 reactions and 204 metabolites.

We applied TFA (164) on a model of central carbon metabolism of mammalian cells in order to characterize thermodynamically feasible intracellular flux states associated with Warburg phenotype based on a pre-selected set of metabolic objectives.

Experimental data from CHO cells during the non-growth phase was used to simulate a healthy/normal phenotype (256) and data collected during the growth phase was used to simulate a cell proliferation phenotype (257, 258). In total 20 (for the normal phenotype) and 21 (for the growth phenotype) exchange fluxes (uptake/secretion) were constrained by experimental data and defined our media. This particular data was chosen due to its similarities in cell culture and the fact that it covered the growth and non-growth phases of the cells.

The reaction thermodynamics constraints added to the FBA problem, even without specific metabolomics measurements, constrained further the solution space.

FVA was used to identify the 102 and 107 bidirectional reactions in the simulated Healthy and Warburg phenotypes, respectively. The addition of thermodynamic constraints further reduced the feasible space of flux solutions and decreased the number of bidirectional reactions by ~25% in both normal and growth types.

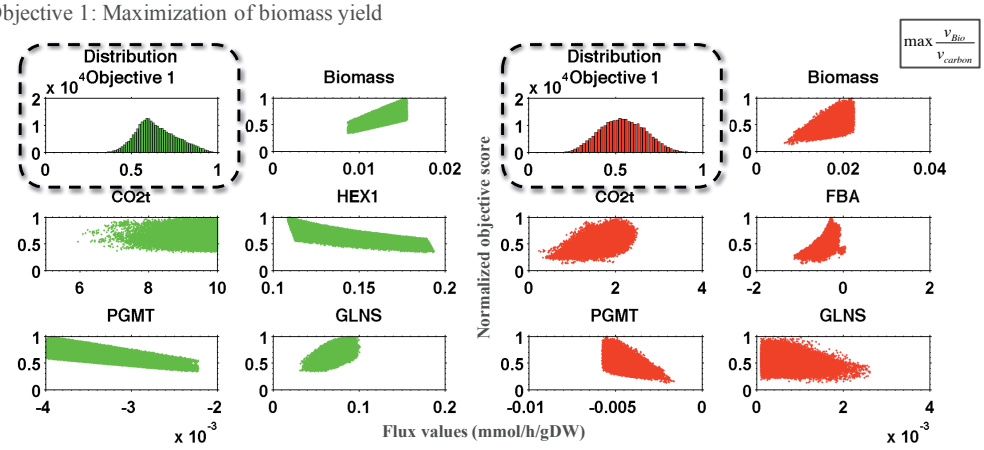


The notion that cellular metabolism, after years of evolution, works towards optimizing a particular objective is widely accepted. However, we have silently defaulted to using maximization of Biomass (or product secretion) as the primary objective. Trying to describe the metabolic flux states based on a single objective can be misleading as it has been shown in (259).

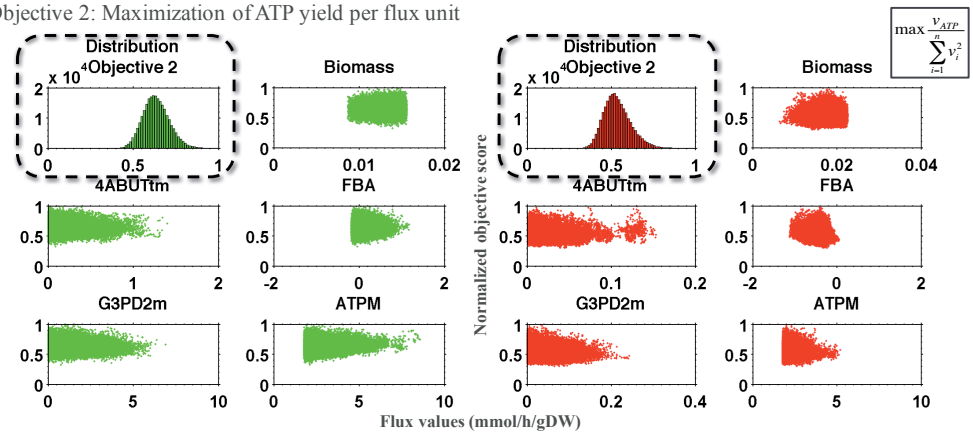
We have used a combination of 3 different metabolic objectives in order to identify changes in central carbon metabolism associated with a switch to an overflow metabolism and higher proliferation state. TFVA was used to identify the thermodynamically permissible range for all the fluxes of the network. The figures below display the correlation between the flux of select reactions in glycolysis and citric acid cycle and the metabolic objectives considered herein, based on 500,000 flux samples. Green and red represent normal and growth phenotypes, respectively.

This type of analysis can help to pinpoint switches in the space of solutions and allow us to navigate a map for metabolic reprogramming possibilities on each side of the switch that are not uniquely dependent on the assumption of an obligatory cellular metabolic objective being attained.

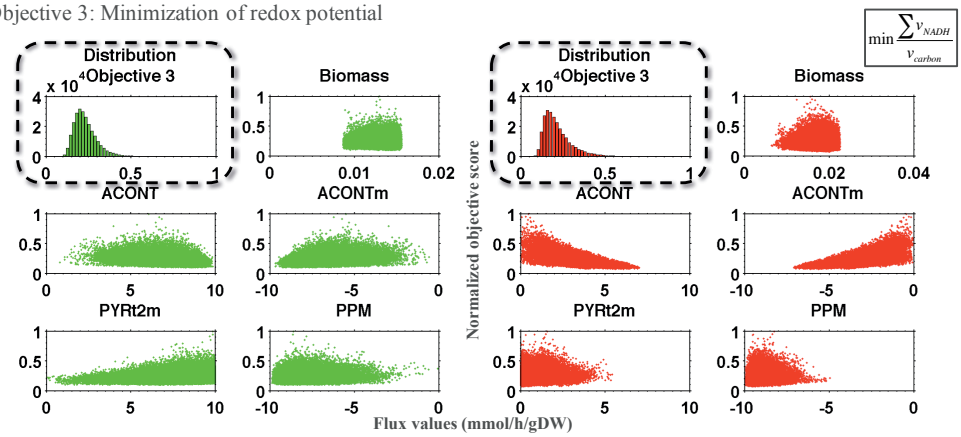
Objective 1: Maximization of biomass yield



Objective 2: Maximization of ATP yield per flux unit



Objective 3: Minimization of redox potential



Appendix B Supplementary Figures

B.1 Supplementary Figures for Chapter 1

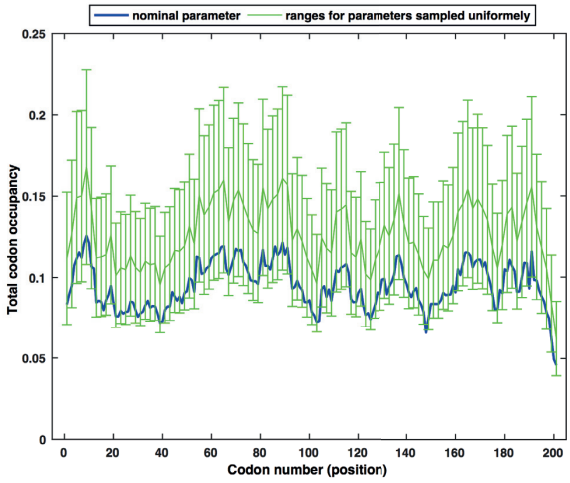


Figure B.1.1 Range of total codon occupancies computed with modified HZ model for *E. coli* gene *yahD* using $2^{12} \cdot (n+2)$ samples of ribosomal kinetic parameters (n is the number of model parameters in Table C.1.1). The blue curve is the total codon occupancies computed using the reference kinetic parameters. Since the translation initiation rate parameter chosen was low, translation of this gene is initiation limited, which justifies the position of the nominal state close to the lower bound of the simulated range.

B.2 Supplementary Figures for Chapter 2

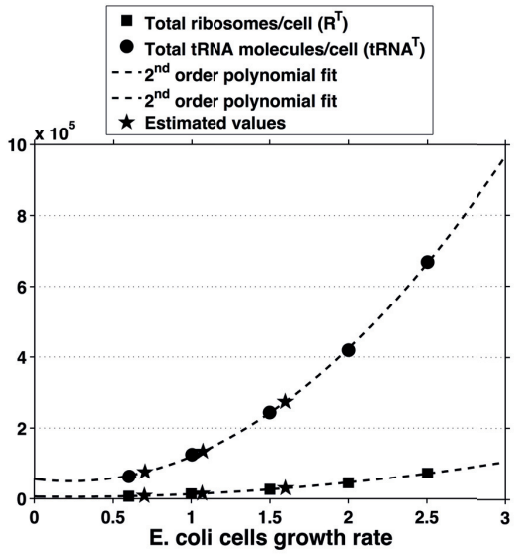


Figure B.2.1 The total number of ribosomes per cell (R^T) (squares) and the total number of tRNA molecules per cell ($tRNA^T$) (circles) were obtained from (58) for the growth rates 0.6, 1.0, 1.5, 2.0, and 2.5 h^{-1}. The respective values for the growth rates 0.4, 0.7, 1.07, and 1.6 h^{-1} for which we know the concentrations of each tRNA species (59) were estimated from a 2^{nd} order polynomial fitting of the data (stars and dashed lines). The $tRNA^T$ for 0.4 h^{-1} was not estimated as the number of molecules per tRNA species was determined in (59).

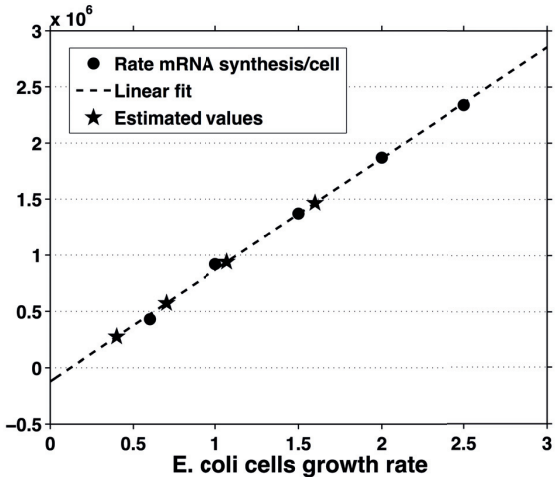


Figure B.2.2 The rate of mRNA synthesis per cell (circles) was obtained from (58) for the growth rates 0.6, 1.0, 1.5, 2.0, and 2.5 h^{-1}. The respective values for the growth rates of interest 0.4, 0.7, 1.07, and 1.6 h^{-1} were estimated from a linear fitting of the data (stars).

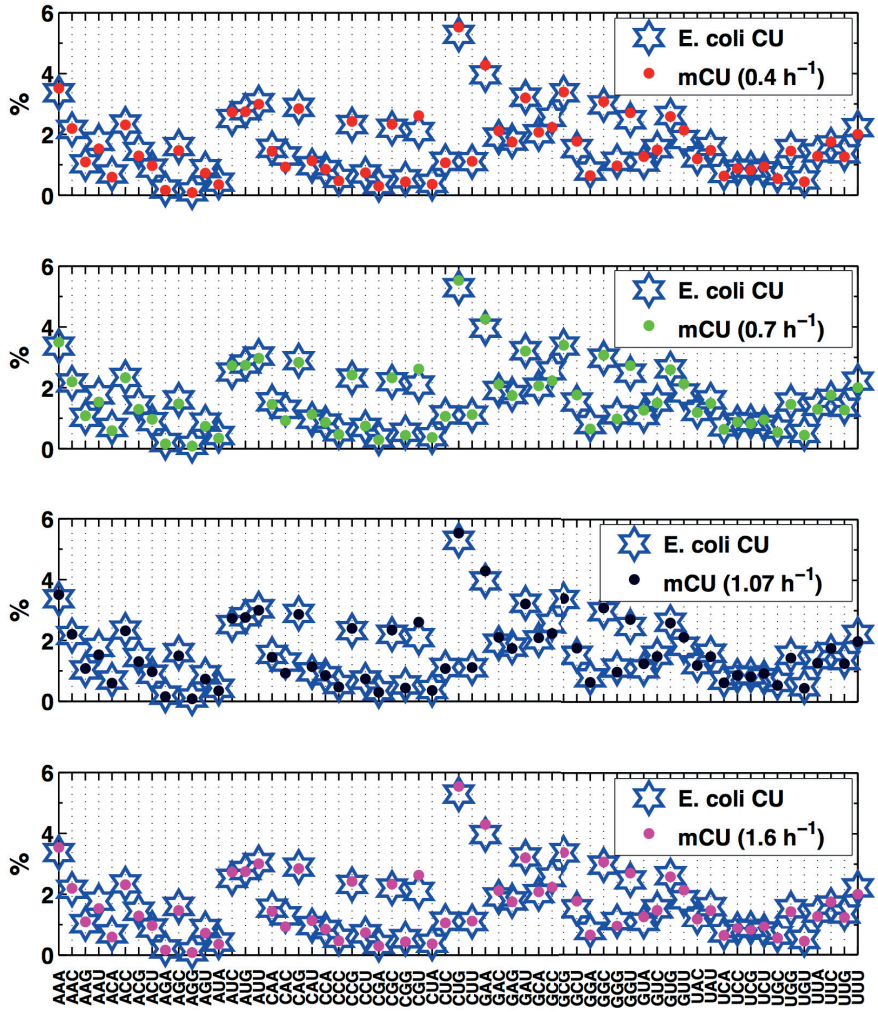


Figure B.2.3 Visual comparison between the *E. coli* codon usage frequency (CU) and the mRNA codon usage frequency (mCU) in the mRNA pools generated for the different growth rates. Since CU is dictated by the organism's genome, it is independent of growth rate.

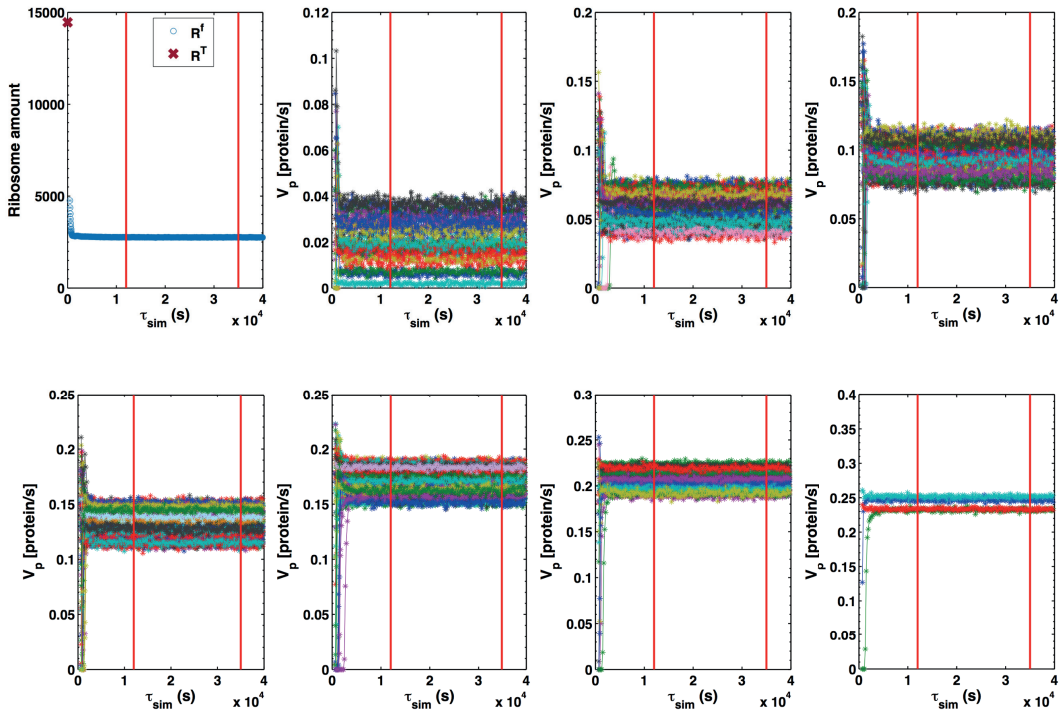


Figure B.2.4 (a) Time evolution of free ribosome amount. At simulation time 0s the number of free ribosomes is equivalent to the total number of ribosomes in the cell. The 'x' in red represents the number of total ribosomes, which remains constant throughout the simulation. (b-h) Time evolution of protein synthesis (V_p) for an *E. coli* cell simulation at a growth rate of 1.07 h^{-1}. Each curve represents the V_p of each of the 289 mRNA species, which are spread over multiple subplots for better visibility, and is averaged over 50 repeated simulations. The red vertical bars indicate the time interval from where data is recorded for the subsequent analysis at steady state. Error bars not included for clarity.

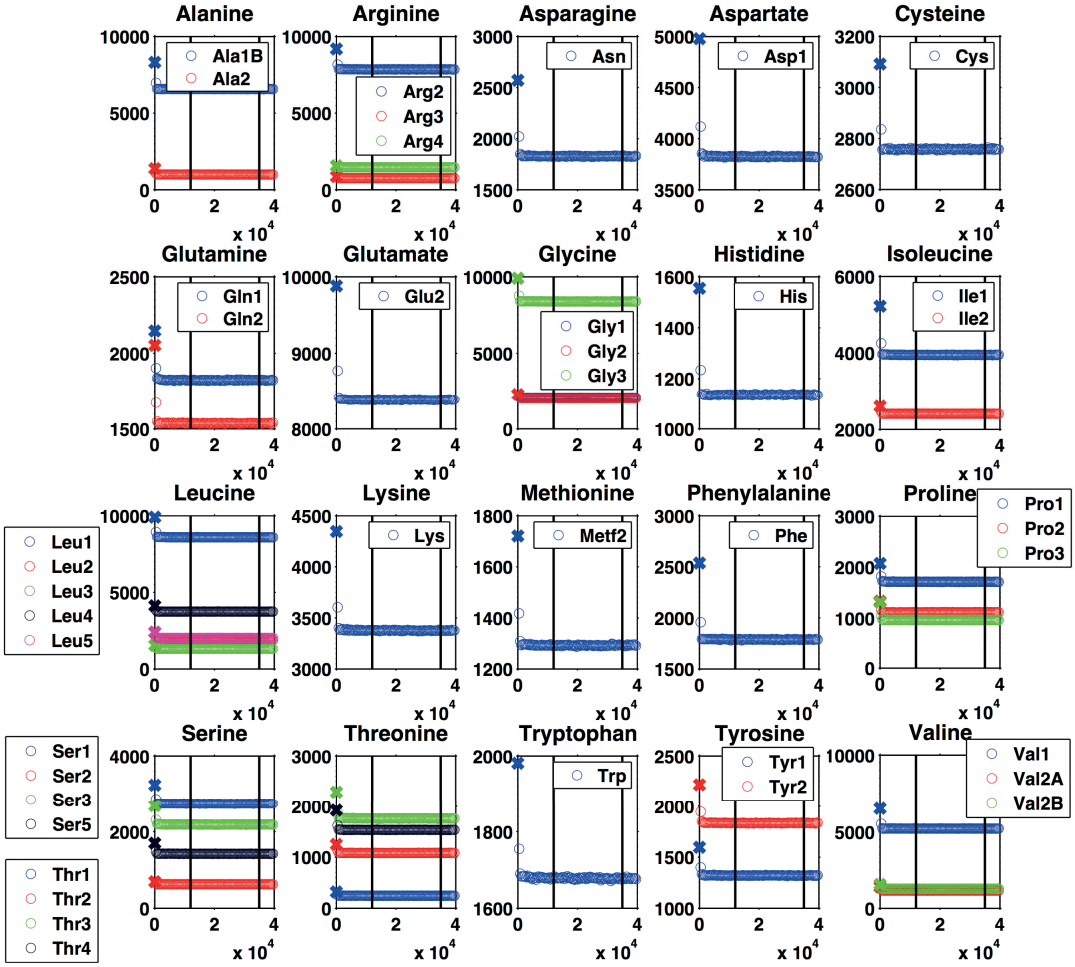


Figure B.2.5 Time evolution of the number of free tRNA molecules of each species for an *E. coli* cell simulation at a growth rate of 1.07 h^{-1} after averaging over 50 repetitions. At simulation time 0s the number of free tRNA molecules of each species is equivalent to the total number of tRNA molecules of each species in the cell. The 'x' represents the number of total tRNA molecules of each species, which remains constant throughout the simulation. Vertical bars indicate the time interval from where data is recorded for the subsequent analysis at steady state. Error bars not included for clarity due to their small size, which can be seen for the same growth rate in Figure B.2.15.

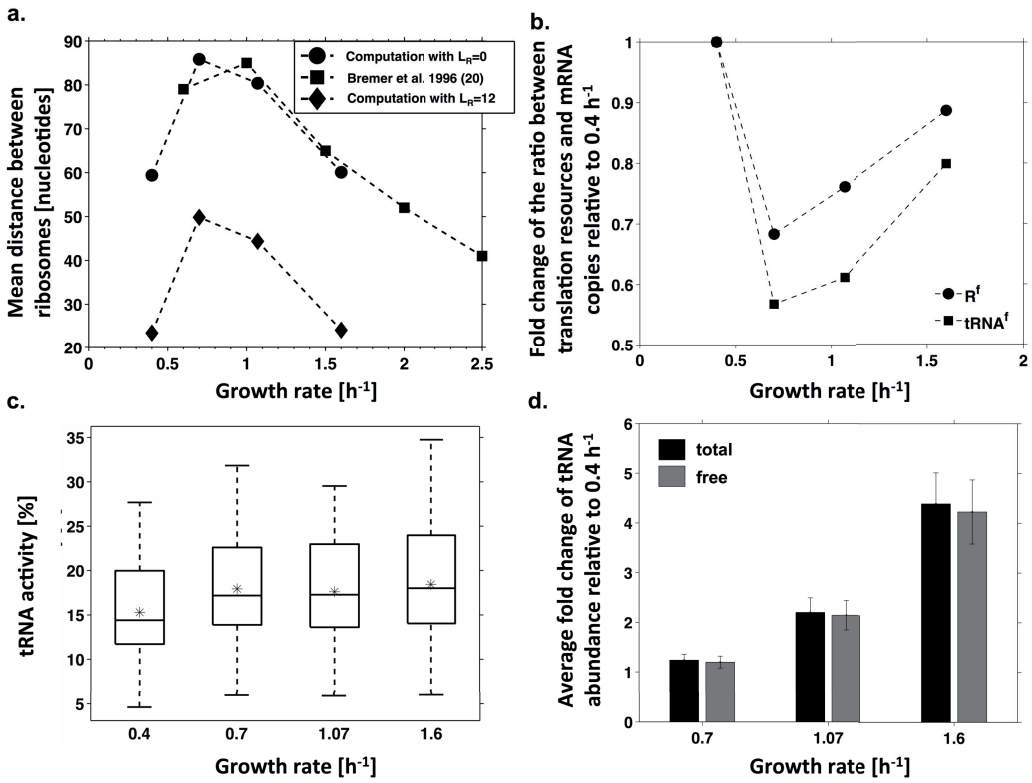


Figure B.2.6 (a) Mean distance between ribosomes in function of growth rate. Results from computations with different assumptions on ribosome length (circles and diamonds) are compared with estimated values from (58) (squares). (b) Fold change of the ratio of R^f/number mRNA copies and $tRNA^f$/number mRNA copies relative to the ratio at 0.4h^{-1} in function of growth rate. (c) Distribution of the activity of each tRNA species for each growth rate. The bar represents the median of the distribution, the ∗ represents the mean of the distribution, the edges of the box are the 25th and 75th percentiles, and the whiskers represent about 99.3% coverage of the data points for data assumed normally distributed. The edges of the whiskers contain the most extreme data point that is not an outlier. (d) Average across all tRNA species of the fold change of the number of total and free tRNA molecules at growth rates [0.7, 1.07, and 1.6] h^{-1}, relative to the number of total and free tRNA molecules at 0.4h^{-1}.

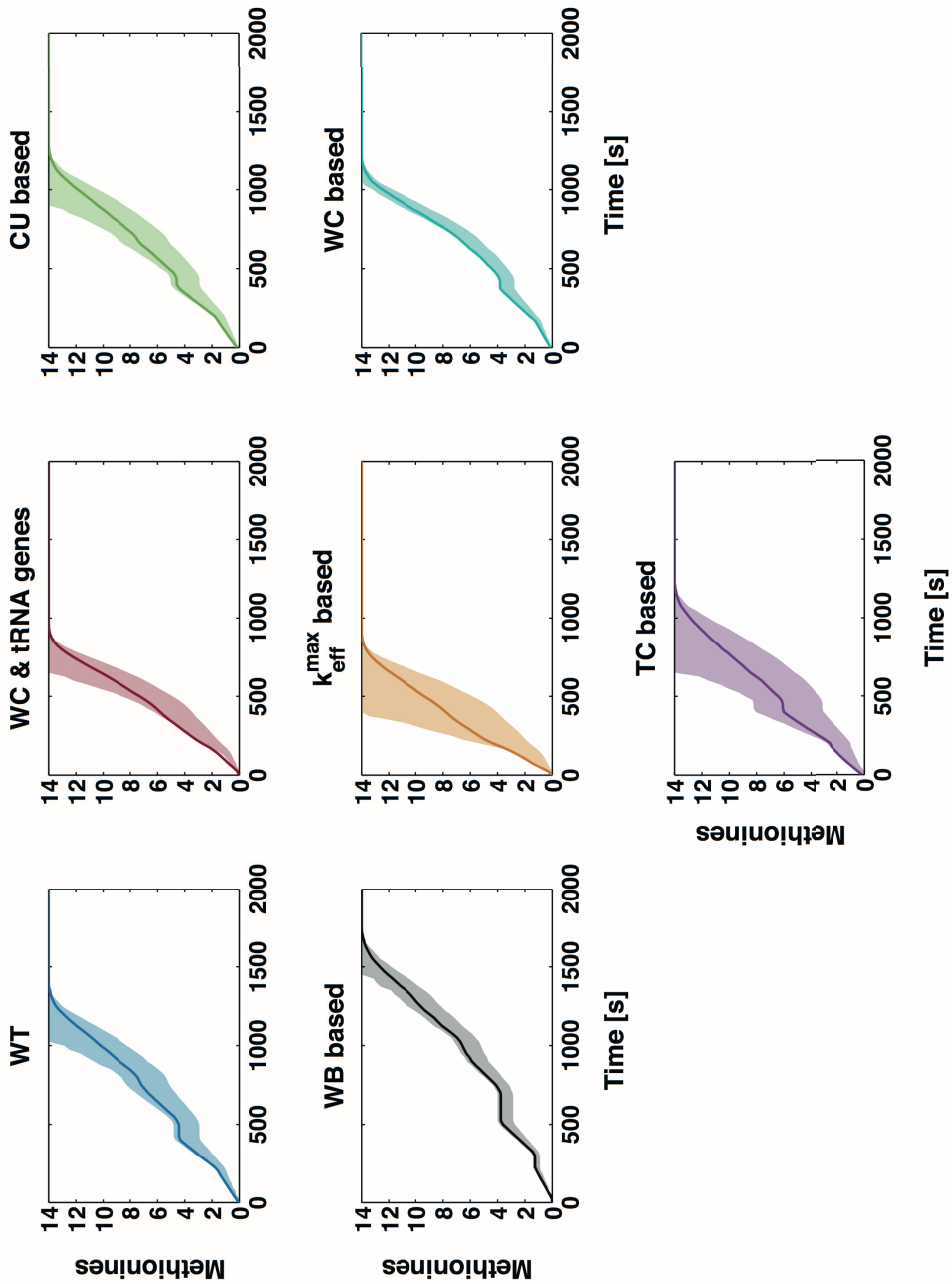


Figure B.2.7 *In-silico pulse chase experiment* curves obtained by averaging the time-evolution of the methionine level in the proteins produced over 4000 simulated cells. Bounds represent the 25th and 75th percentile of the distribution.

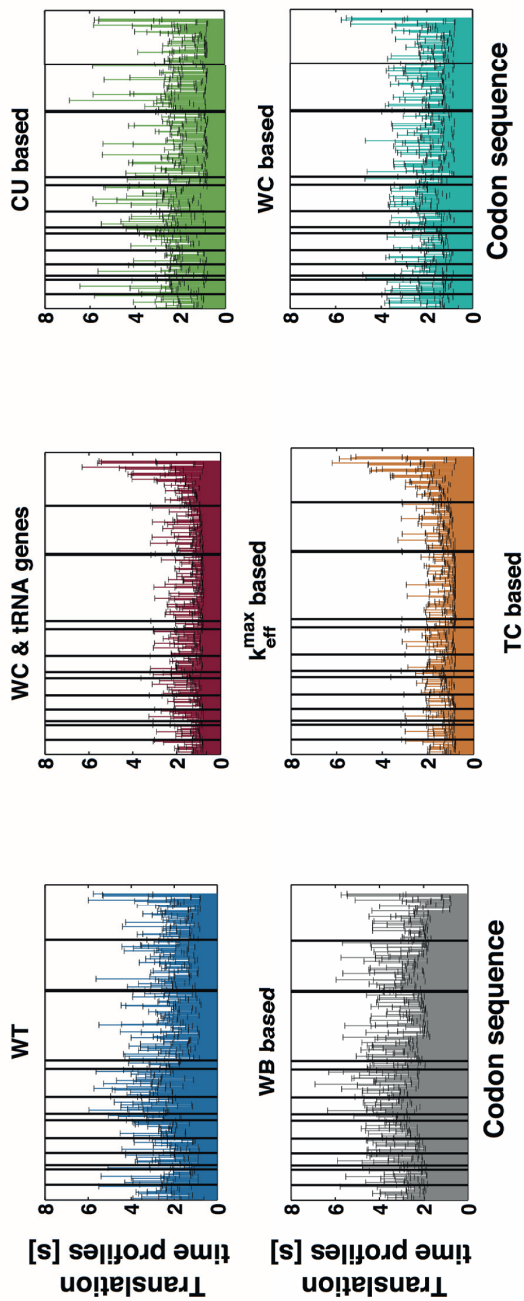


Figure B.2.8 Translation time profiles for each of the seven Luciferase transcripts studied. Black vertical bars indicate the position of the methionine encoding codons in the sequences, which is the same for all. Translation profiles of k_{eff}^{max} based and WC & tRNA genes based transcripts are faster since all codons have been substituted by synonymous codons that maximized the number of WC interactions with cognate tRNAs. Note that since transcript level in bacteria correlates well with gene level, transcripts built based on WC & tRNA genes design have their codons substituted by codons that have also the highest number of WC tRNA cognate pairs. Since these two sequences are the best at optimizing codon elongation rate, with k_{eff}^{max} based being the most optimal, it is then observed a queuing of the ribosomes starting from the termination codon as this step is now the most limiting during translation of the optimized sequences.

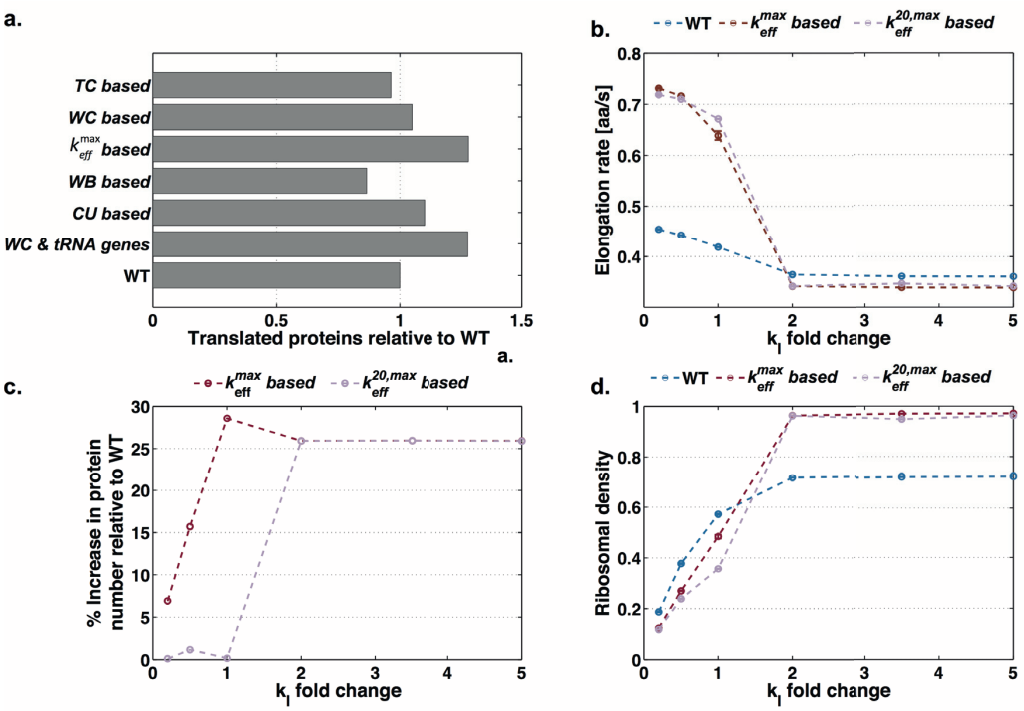


Figure B.2.9 (a) Fold change in protein translation from all transcripts relative to WT. (b) Elongation rate in function of translation initiation rate constant for WT, k_{eff}^{max} *based* and $k_{eff}^{20,max}$ *based*. k_{eff}^{max} *based* has the same codon sequence except for the 20 first codons that match the ones in the WT ($k_{eff}^{20,max}$ *based*). Since the value of (k_{I}) is set equal between the transcripts the change in initiation rate based on synonymous codon substitution of the first codons results from the variation on ribosome binding space at the beginning of the sequence. For k_{I} fold change 1 the elongation rate remains practically unchanged for these two transcripts, confirming that the different pulse-chase curves between WT and k_{eff}^{max} *based* are indeed a result of changes on elongation rate rather than initiation. (c) Increase in the number of proteins translated from k_{eff}^{max} *based* and $k_{eff}^{20,max}$ *based* transcripts relative to WT. (d) Ribosomal density in function of translation initiation rate constant for WT, k_{eff}^{max} *based* and $k_{eff}^{20,max}$ *based*. There is no increase in protein synthesis of $k_{eff}^{20,max}$ *based* with respect to the WT when compared to k_{eff}^{max} *based* as less ribosome bind to $k_{eff}^{20,max}$ based (see c-d at k_{I} fold change 1). The increase of the transcript translation initiation rate constant (k_{I}) has the potential to increase protein synthesis until the sequence is fully saturated with ribosomes, whereas elongation rate decreases to a minimum value as a result of the high number of interactions between queuing ribosomes. Even though a transcript is optimized for elongation, we note that the translation initiation rate, which is dictated by the beginning of the transcript's coding region and the steady state R^f of the host cell, has a major impact on the gain in protein production with respect to the WT in its rate limiting regime.

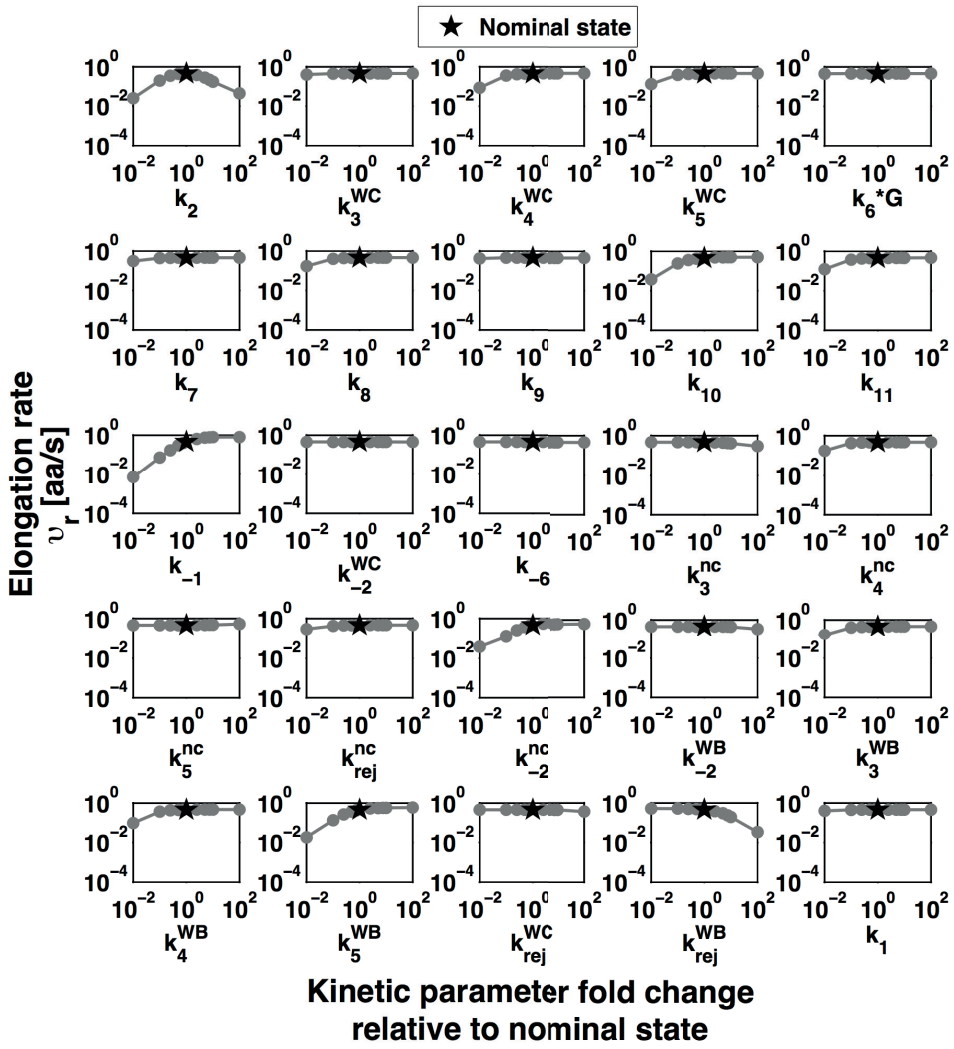
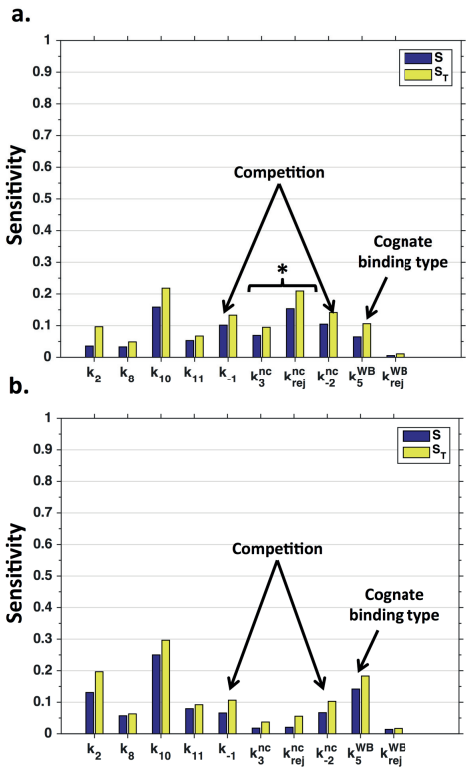


Figure B.2.10 Elongation rate (υ_r) of an mRNA species in function of the change of each of the ribosome kinetic rate constants in the range of two orders of magnitude with respect to their nominal (experimental) values.



c.

	WC	*WB*	*nc*	*non*
$\overline{\Delta t}_{bi}^{A-OFF}$ [$\times 10^{-4}$ s]	2.4 (0.002)	2.9 (0.006)	10.5 (0.003)	5.9 (0.0007)
N_{bi}^{A-OFF}	191652	216501	9384712	31578929
% *	40.5	41.16	99.76	100
$\overline{\Delta t}_{bi}^{A-PROOF}$ [s]	$6{\cdot}10^{-3}$ ($6{\cdot}10^{-5}$)	0.09 ($8{\cdot}10^{-5}$)	0.16 ($7{\cdot}10^{-4}$)	-
$N_{bi}^{A-PROOF}$	933	112851	21259	-
% *	0.2	21.45	0.23	-
$\overline{\Delta t}_{bi}^{P-ON}$ [s]	0.30 (0.0007)	0.39 (0.0005)	0.46 (0.005)	-
$\overline{\Delta t}_{bi}^{E-OFF}$ [s]	0.74 (0.04)	0.83 (0.03)	0.89 (0.03)	-
$N_{bi}^{P-ON/E-OFF}$	280582	196640	952	-
% *	59.3	37.38	0.01	-
t_{codon}^{bi} [s/aa]	0.44	0.36	0.03	0.04

% total binding events	*Cognate (WC & WB)*	*Competition (nc & non)*
	2.4	97.6

* Fraction of *bi* events (*WC*, *WB*, *nc*, *non*) per decoding stage (A-site OFF, A-site PROOF and P-site ON/E-site OFF).

(values) are standard deviations.

Figure B.2.11 Sensitivity of ribosome kinetic parameters (a) measured at *in vitro* conditions at 37ºC and (b) deduced in (94) for *in vivo* conditions. Each rate constant of the *WB* kinetics (only available at 20ºC) was scaled by the kinetics of the *nc* pathway for these two conditions by maintaining the ratios between *WB* and *nc* kinetics at 20ºC. The values for translocation kinetics were maintained at 25ºC, as we have no information for their changes, which increased their sensitivity compared to 20ºC. However, as these values are expected to increase with temperature (94) to match the observed elongation rates, this limitative effect on elongation rate will also decrease. Parameters * become more influential at 37ºC than at 20-25ºC and *in vivo* because the net rate constants of the near-cognate pathway until GTP hydrolysis are higher at 37ºC. In the overall tRNA competition becomes less important that the cognate binding type for conditions closer to *in vivo*. (c) Statistics on mean ribosome occupancy time lags and total number of events per decoding stage and binding type for the simulations with ribosome kinetic parameters deduced in (94) for *in vivo* conditions.

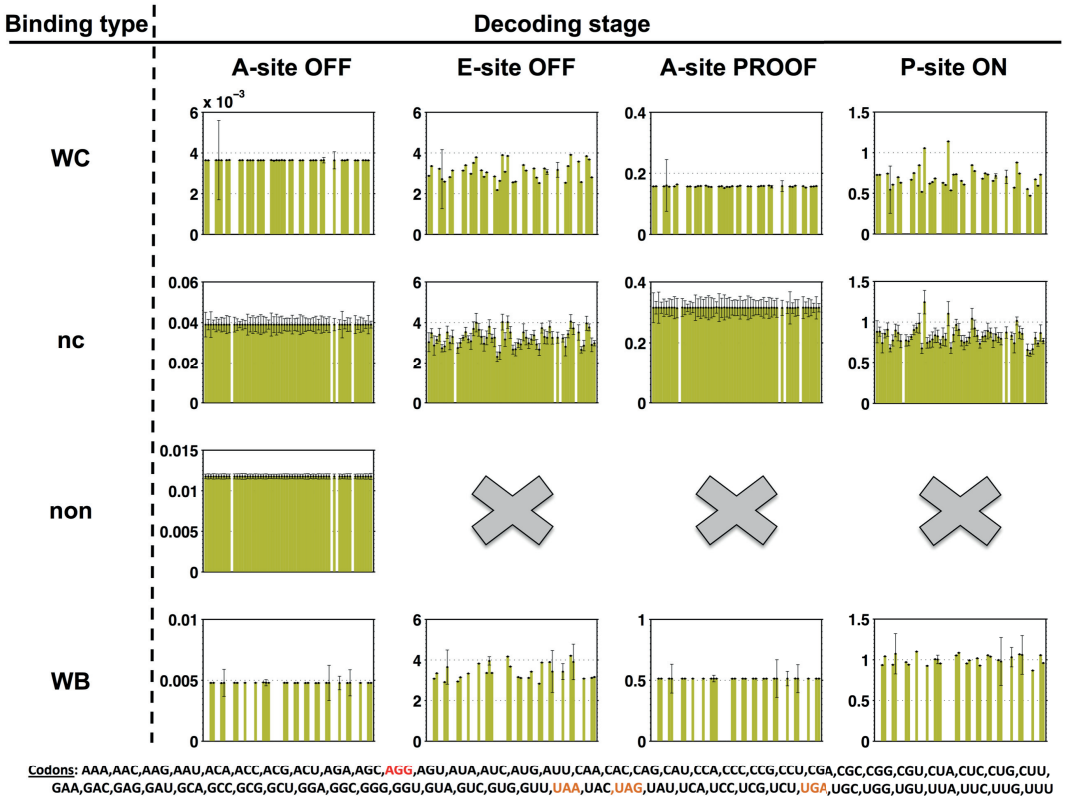


Figure B.2.12 *Mean ribosome occupancy time lags* $\left(\overline{\Delta t}_{bi}^{ds}\big|_{j}\right)$ *per decoding stage* ds = (A-site OFF, A-site PROOF, P-site ON, E-site OFF), *binding interaction* bi = (*WC, WB, nc, non*) and per codon species j. The mRNA species used for the estimations was the WT Luciferase. For each plot the order of the codon identities is the one indicated below. Codons highlighted in orange are stop codons and statistics are not computed for these. The codon highlighted in red is not present in the WT Luciferase sequence. Time lags are very homogeneous among codons as they depend only on the intrinsic ribosome kinetics, which is the same for all codons. Only for the decoding stages that include ribosome translocation (P-site ON and E-site OFF) the time lags become more heterogeneous as they become dependent on the blocking from downstream ribosomes.

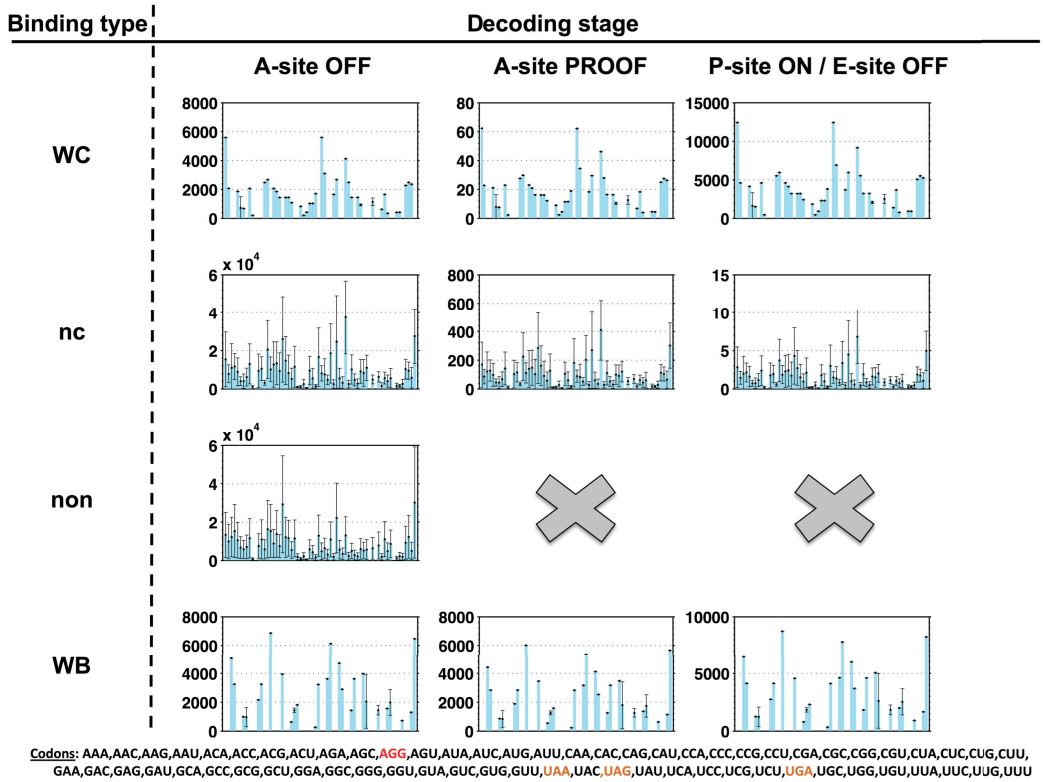


Figure B.2.13 *Total number of events* $\left(N_{bi}^{ds}\big|_{j}\right)$ *per decoding stage* ds = (A-site OFF, A-site PROOF, P-site ON, E-site OFF), *binding interaction* bi = (*WC, WB, nc, non*) and per codon species j. The mRNA species used for the estimations was the WT Luciferase. For each plot the order of the codon identities is the one indicated below. Codons highlighted in orange are stop codons and statistics are not computed for these. The codon highlighted in red is not present in the WT Luciferase sequence. The number of events varies among codons because they are dependent on both mCU and the free tRNA abundances.

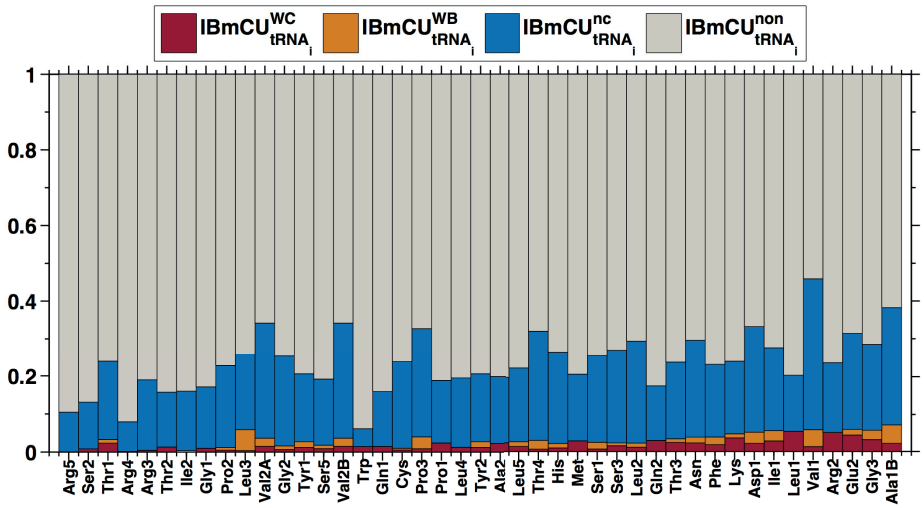


Figure B.2.14 Interaction-based mRNA codon usage frequency (*IBmCU*) displayed for each tRNA species. The x-axis is arranged in increasing order of the tRNA activity observed in Fig. 5 of main text. mCU is grouped into five *interaction-based mRNA codon usage frequency* groups for each tRNA species i ($IBmCU^{cogn}_{tRNA_i}$, $IBmCU^{WC}_{tRNA_i}$, $IBmCU^{WB}_{tRNA_i}$, $IBmCU^{nc}_{tRNA_i}$, $IBmCU^{non}_{tRNA_i}$) (Eq. 2.1 main text).

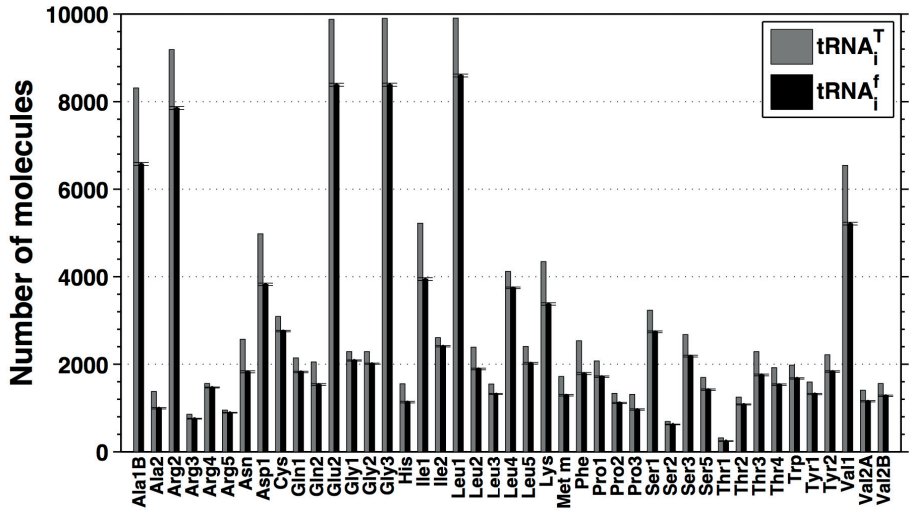


Figure B.2.15 Comparison of total number of tRNA molecules ($tRNA_i^T$) and the number of free tRNA molecule at steady state ($tRNA_i^f$) for each species i at 1.07 h^{-1}.

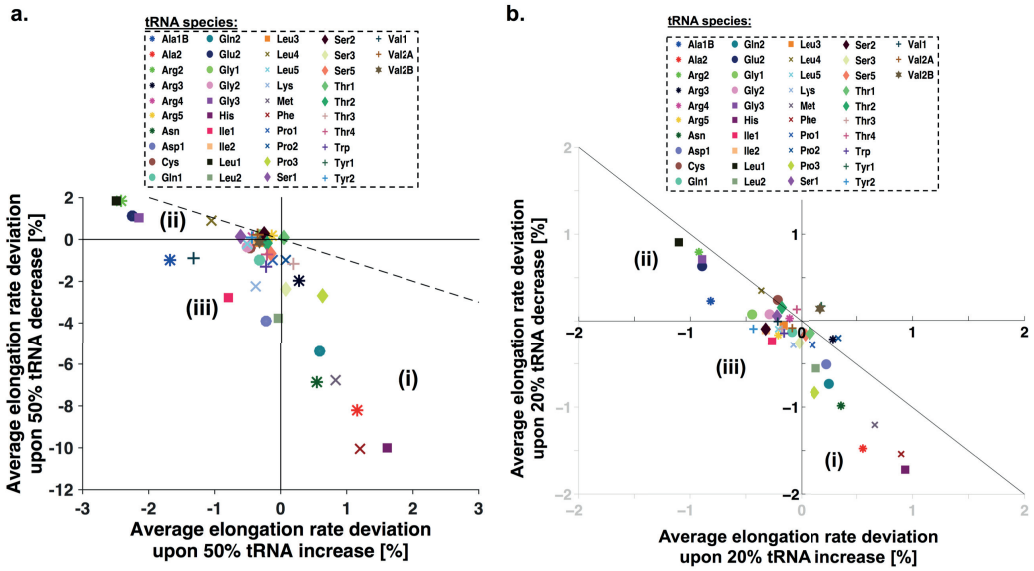


Figure B.2.16 Relative deviation of average elongation rate from all mRNA species in the cell at $1.07h^{-1}$ upon (a) 50% and (b) 20% increase or decrease of each tRNA abundance. Worthy to mention, with respect to the results obtained for changes of 50%, is the reassignment of Ala1B and Val1 to group (ii) and Asp1 and Leu2 to group (i), whereas for the 50% change case they all belonged to group (iii). Although the effect on Ala1B and Val1 surplus remains the same, starvation of only 20% of these tRNAs is still not sufficient to cause a negative impact on their cognate codons. For Asp1 and Leu2 surplus of 20% is still not sufficient to negatively affect the elongation rates by means of increasing competition on their near- and non-cognate codons.

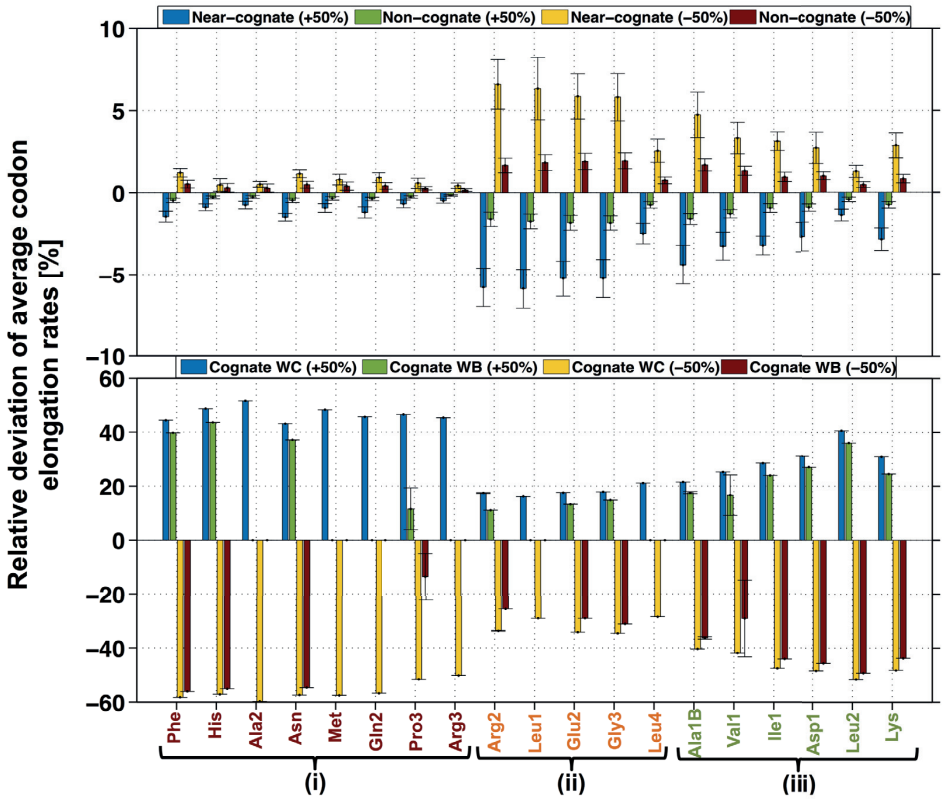


Figure B.2.17 Relative deviation of average codon elongation rate (computed with k_{eff}) from all codons that are cognate *WC*, cognate *WB*, near-cognate and non-cognate to the tRNA species whose concentration is changed by ±50%. Red, orange and green text colors differentiate between tRNA species that belong to regimes (i), (ii), and (iii), respectively. (i) tRNA species whose surplus or starvation contribute to the biggest increase or decrease, respectively, of the average codon elongation rate of their cognate codons. These tRNAs are among the ones whose cognate (specially *WB* type) codons have very slow codon elongation rates and appear frequently in the mRNA sequences (Figues B.2.18 and B.2.19). (ii) tRNAs whose starvation or surplus contribute to an increase or decrease, respectively, of the average codon elongation rate of their near- and non-cognate codons. These tRNAs are among the species in the cell that are present in higher abundances (Figure B.2.15) and, as a consequence, have the highest cognate codon elongation rates. The surplus of these tRNAs acts on the system by decreasing the mean elongation rate due to the prominent tRNA competition they provide to their near- and non-cognate codons. Under starvation the effect is the inverse since the level of competition is decreased on the near- and non-cognate codons. These tRNAs are so abundant that even when their number is decreased their cognate codons do not limit translation. Leu4 is an exception in this group and its qualification results from the fact that it belongs to the top 10 tRNAs species with highest combined $IBmCU^{nc}_{Leu4} + IBmCU^{non}_{Leu4}$ and from these 10 it is the one with the highest free tRNA abundance, such that its surplus or starvation is able to moderately affect the system. (iii) tRNA that have an effect similar to (ii) under surplus due to the increase in competition resulting from their high abundances (but less than (ii)) or high $IBmCU^{nc}_{tRNA_i} + IBmCU^{non}_{tRNA_i}$, as in the case of Leu2, however, under starvation their cognate codon elongation rates are negatively affected by the high $IBmCU^{cogn}_{tRNA_i}$ demanding free tRNA.

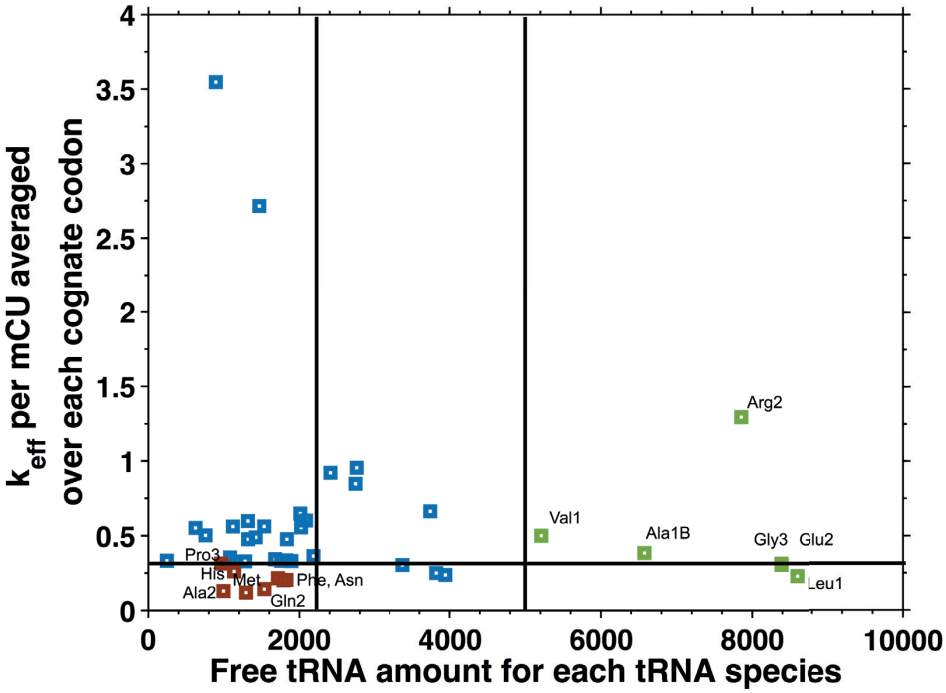


Figure B.2.18 For each tRNA species in the cell, the k_{eff} values of all its cognate codons (*WC* and/or *WB*) are divided by their correspondent mCU quantities and averaged together. The reported quantity is obtained for each tRNA species and plotted against its corresponding free amount in the cell. Red-labeled tRNA species belong to regime (i) and the green-labeled ones belong to regime (ii).

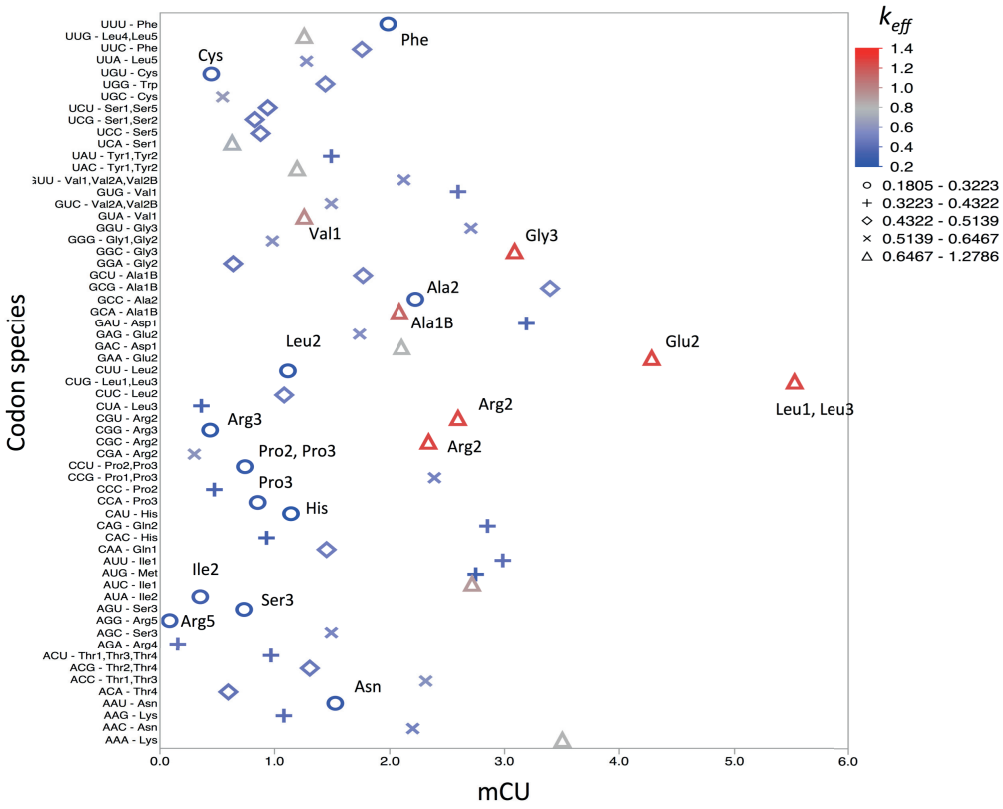


Figure B.2.19 mRNA codon elongation rate (mCU) presented per codon species and tRNA isoacceptor. Color code represents the values of the codon elongation rate for each codon species (k_{eff}).

B.3 Supplementary Figures for Chapter 3

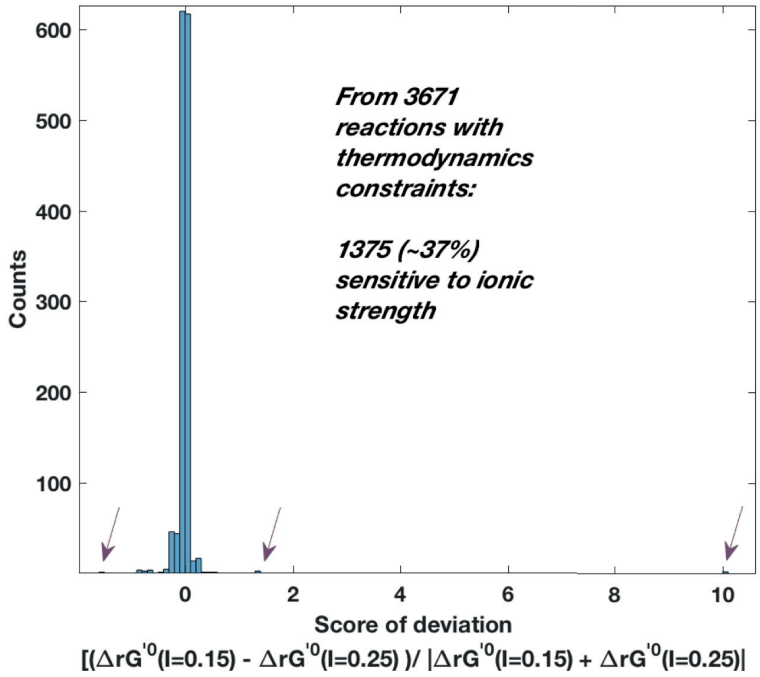


Figure B.3.1 Distribution of the deviation scores computed as $\frac{\left(\Delta_r G'^{\circ}(I=0.15)-\Delta_r G'^{\circ}(I=0.25)\right)}{\left|\Delta_r G'^{\circ}(I=0.15)+\Delta_r G'^{\circ}(I=0.25)\right|}$. The distribution pertains to only ~37% of the reactions with estimated $\Delta_r G'^{\circ}$, to exclude the zero scores of reactions not affected by ionic strength. The arrows indicate the 5 reactions that had the biggest deviation score and presented a switch of sign in the estimated $\Delta_r G'^{\circ}$ at the two ionic strength conditions.

B.4 Supplementary Figures for Chapter 4

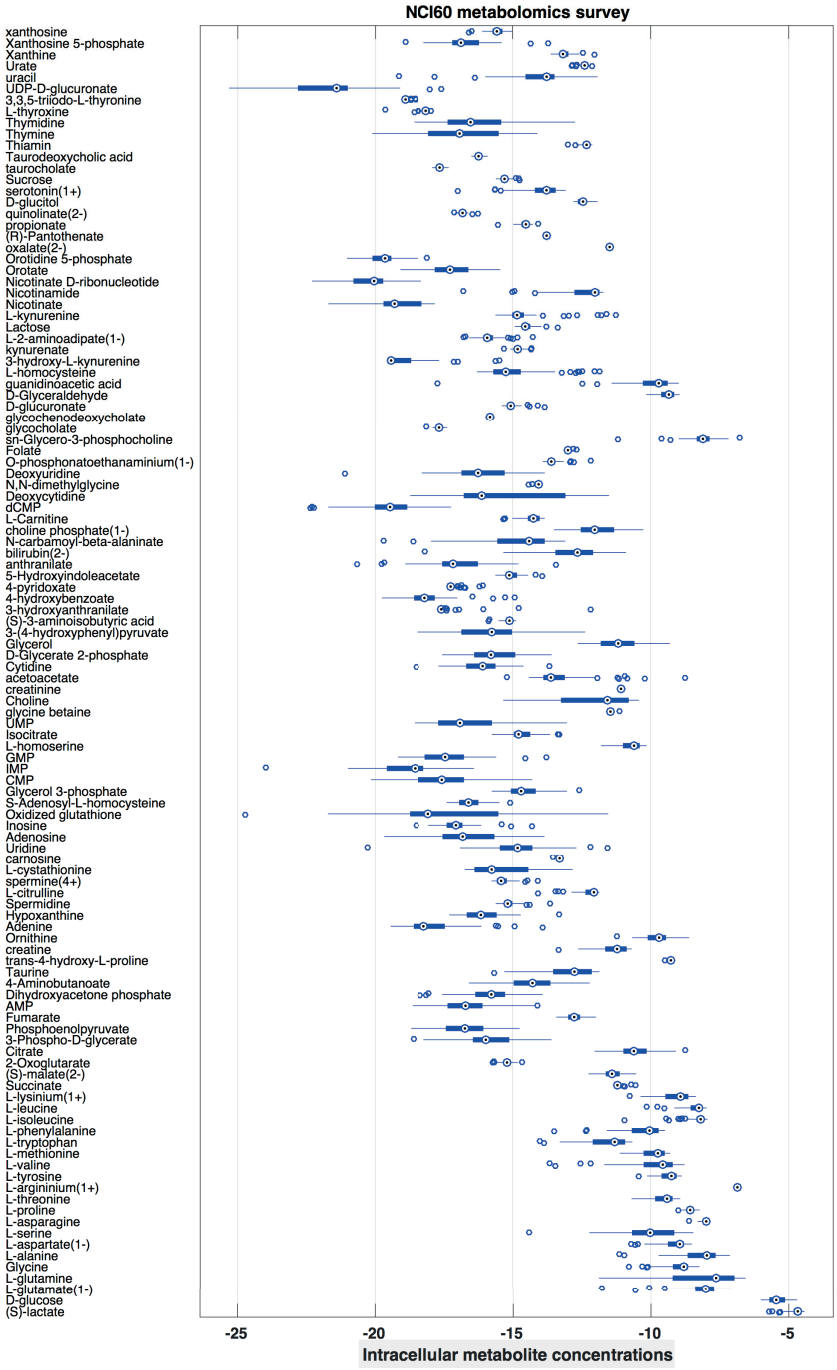


Figure B.4.1. Distribution of intracellular metabolite levels measured for the 60 cancer cell lines in NCI60 panel (217). Values are plotted as log transformed.

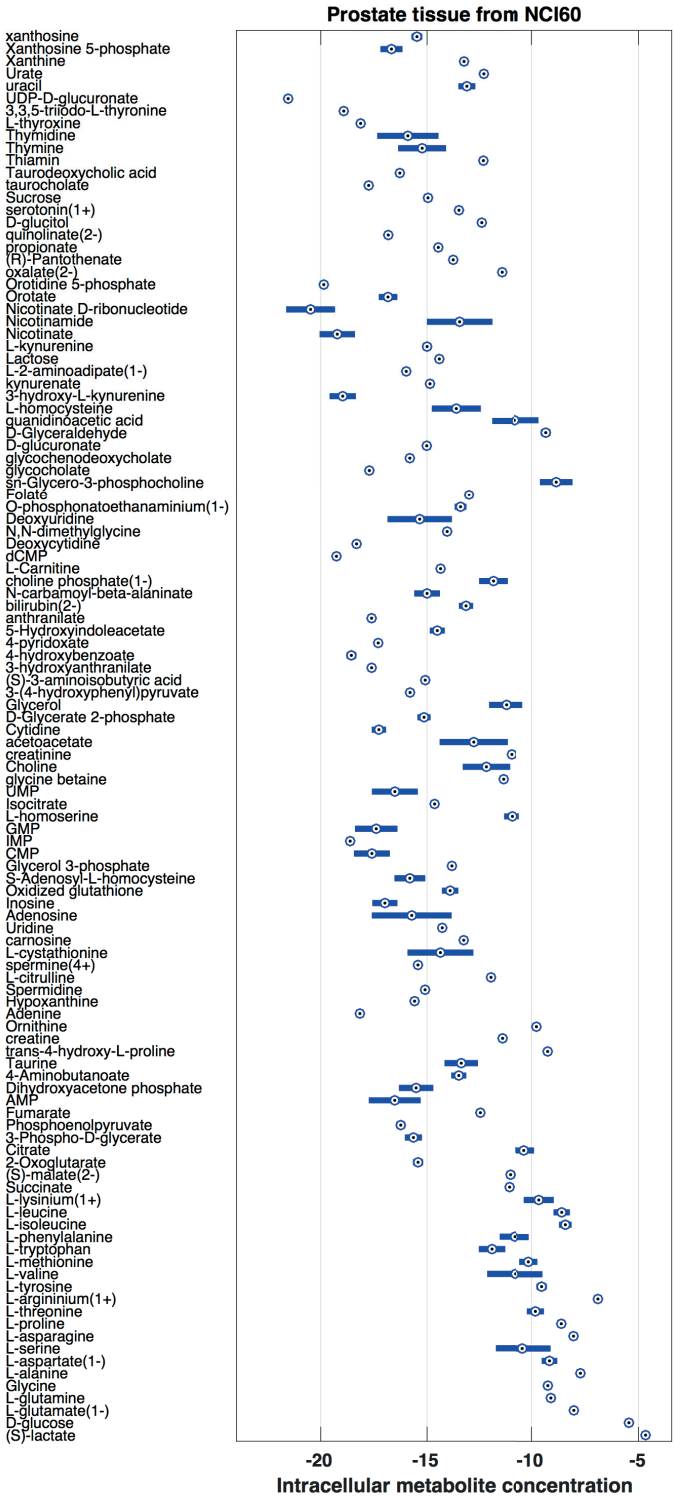


Figure B.4.2 Distribution of intracellular metabolite levels measured for the 2 prostate cancer cell lines in NCI60 panel (217). Values are plotted as log transformed.

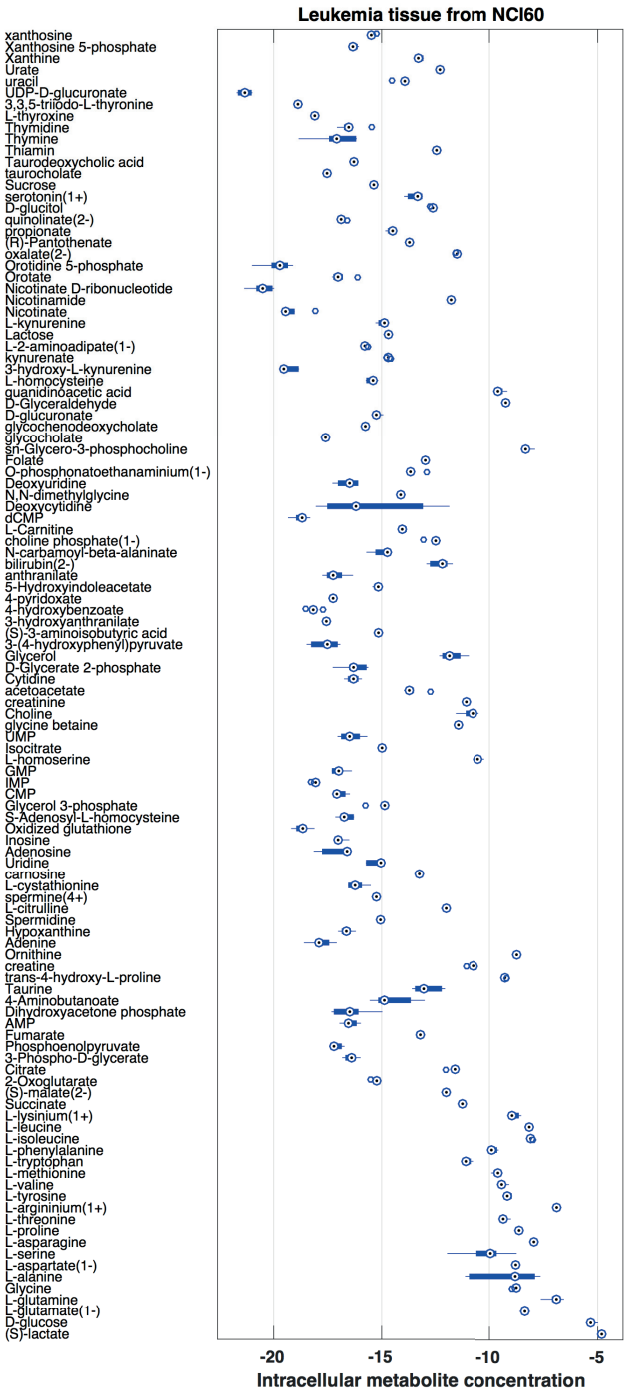


Figure B.4.3 Distribution of intracellular metabolite levels measured for the 6 leukemia cancer cell lines in NCI60 panel (217). Values are plotted as log transformed.

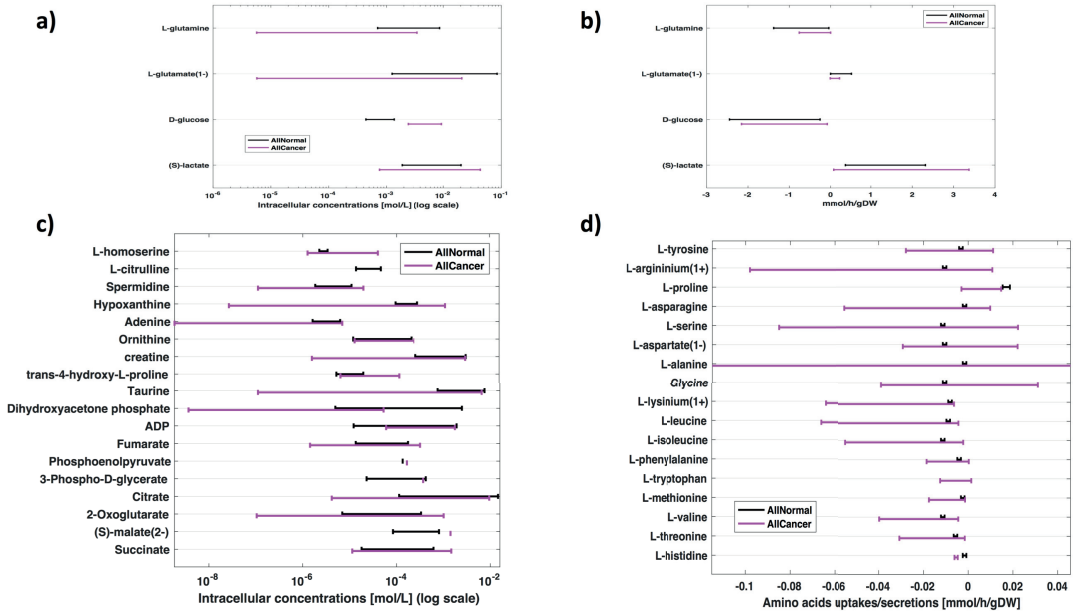


Figure B.4.4 Comparison metabolomics and fluxomics datasets merged to produced normal vs cancer phenotypes. a) Glucose intracellular concentration is higher in cancer cells, whereas glutamine and glutamate present lower range values indicating less accumulation of these metabolites. B) Lactate secretion presents more variability in cancer phenotype than in normal. c) Phosphoelnolpyruvate and (S)-malate(2-), which are key metabolites in pyruvate are lower in cancer metabolism and TCA cycle show differences in cancer vs normal cells. d) The amino acid uptake and secretion rates are much more constrained than the ones measured in cancer, which as expected have much higher variability. We note that we are using for the numeral phenotype measurements of lung fibroblasts during growth, which may present characteristics of an overflow metabolism to support biosynthesis.

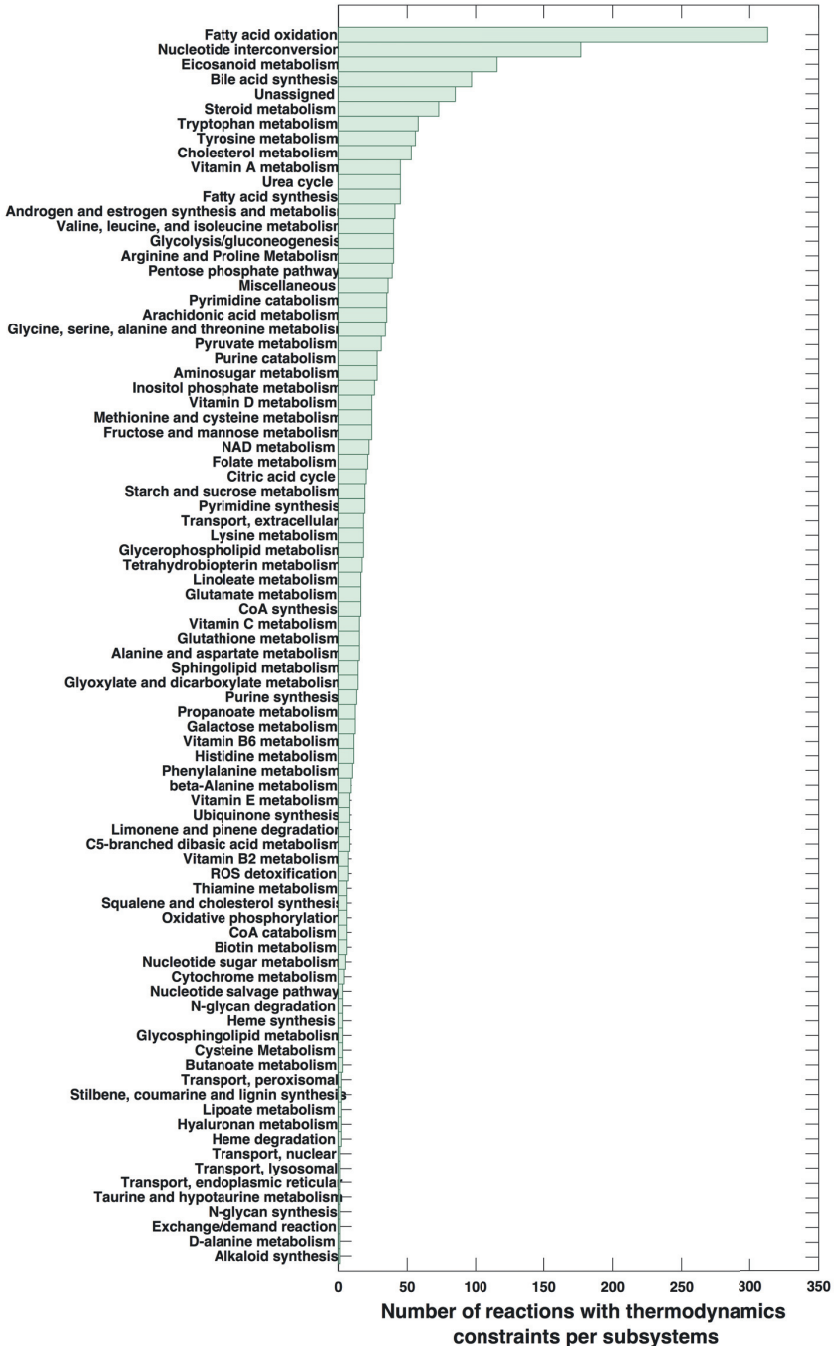


Figure B.4.5 Number of metabolic reactions per subsystem (excluding transport reactions) in Recon 2 v4 that have reaction thermodynamics constraints.

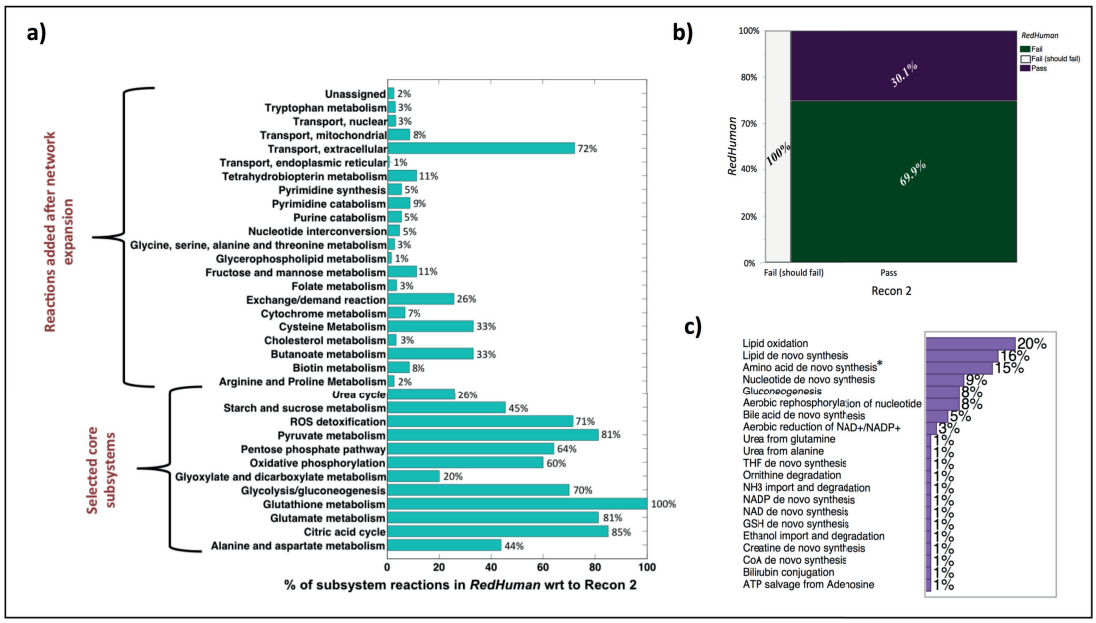
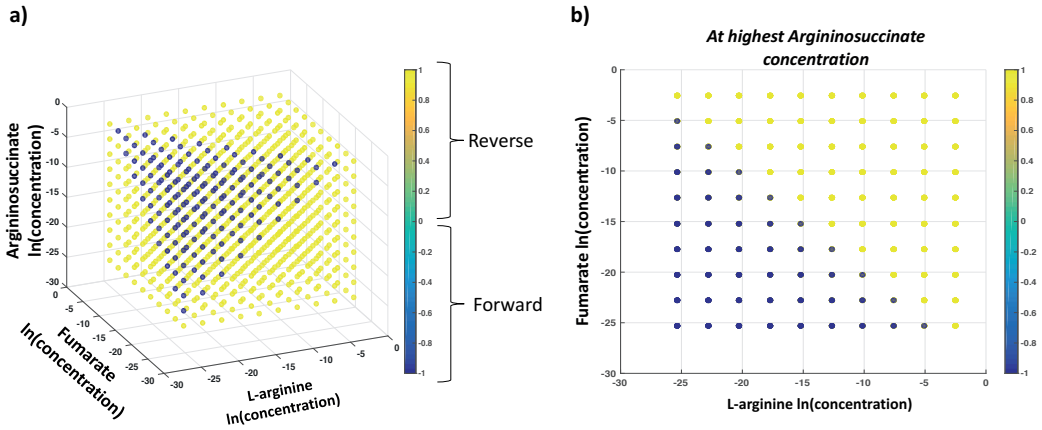


Figure B.4.6 Survey of first human reduction attempt departing from Recon 2 v4 and using fluxomics data merged from CH and NH datasets. The purpose was to test the reduction workflow for the derivation of a reduced model (*RedHuman*) that encompassed normal and cancer phenotypes. a) Percent of reactions preserved from Recon 2 in *RedHuman* within the core network selected at stage 1 and after expansion at stage 2. b) Quantification of human/mammalian related metabolic tasks passed or failed by *RedHuman* with respect to Recon 2 tested metabolic capabilities. c) Classification of failed metabolic tasks that should pass ranked per network subsystem. The bulk of these failed metabolic tasks correspond to parts of the network that we have not selected as core subsystems and hence were not preserved in the resulting *RedHuman*, either because they were beyond the selected threshold for network expansion and out of the network region of interest for our studies, or for carrying no flux in their reactions as they were not connected to the production of biomass building blocks, or because these pathways are non-explicitly preserved in the lumped reactions. We note that this analysis can be helpful to identify problems that may occur due to overlooking certain steps during lumpGEM. For instance, failure in passing metabolic tasks can be related with an overlook on the medium allowed for the cells when producing the lump reactions for each biomass building block. Here the uptake of the non-essential amino acids was not blocked and the shortest route was for the cell to immediately uptake those amino acids to go directly into the biomass, hence rendering the amino acid biosynthesis pathways useless.



ASS1: ATP [cytosol] + L-aspartate [cytosol]+ L-citrulline [cytosol] ⟷ AMP [cytosol] + Ppi [cytosol] + argininosuccinate [cytosol] [ΔrG'°<0]
ASL: argininosuccinate [cytosol] ⟷ L-arginine [cytosol] + fumarate [cytosol] [ΔrG'°>0]

Figure B.4.7 a) Combination of intracellular concentration values for argininosuccinate, fumarate and L-arginine and the respective directionality of ASL reaction. b) Visualization of the plane for which the argininosuccinate concentration is the highest. Yellow box: the two reactions in arginine biosynthesis and their reaction directionality for the intracellular concentrations in both our cancer and normal physiology datasets.

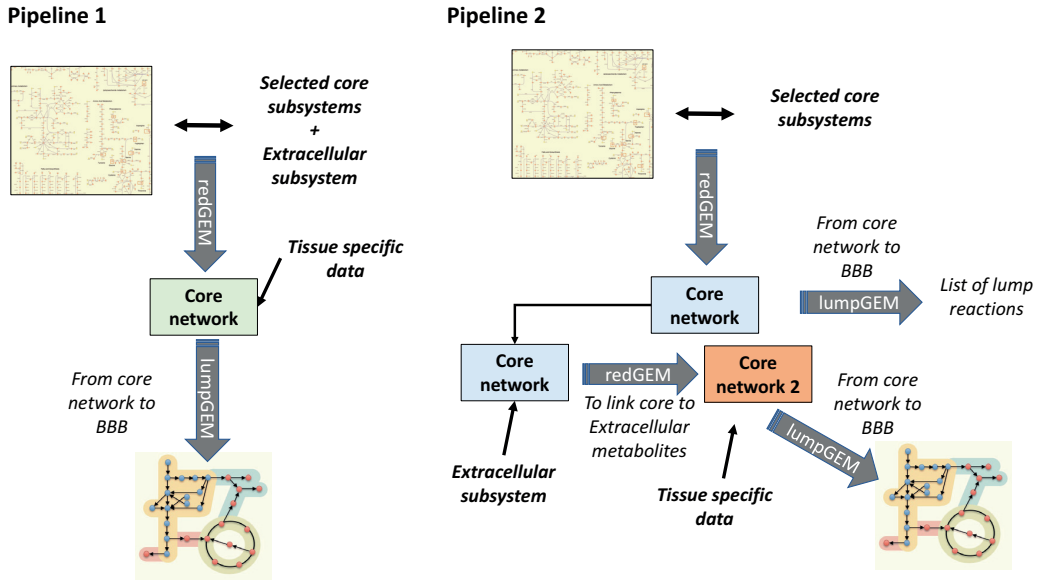


Figure B.4.8 The reduction procedure in redGEM and lumpGEM (205, 206) can be adapted be performed in stages that allow to generate lumps according to the tissue specific data that is integrated after generating the core network, thus allowing for direct comparision of intrisinc network heterogenetiy among different tissues and conditons.

Appendix C Supplementary Tables

C.1 Supplementary Tables for Chapter 1

Table C.1.1 Rate constants for the ribosomal kinetic pathway during translation elongation obtained from experimental sources.

Rate constants	Definition	Cognate (c)	Near-cognate (nc)	Non-cognate (non)
$k_1(\mu M^{-1}s^{-1})$	Initial binding	140 *,†	140 *,†	140 *,†
$k_{-1}(s^{-1})$	Reverse initial binding	85 *,†	85 *,†	85 *,†
$k_2(s^{-1})$	Codon reading	190 *	190 *	—
$k_{-2}(s^{-1})$	Reverse codon reading	0.2 *	80 *	—
$k_3(s^{-1})$	GTPase activation and GTP hydrolysis	260 *	0.4 *	—
$k_4(s^{-1})$	Pi release and EF-Tu rearrangement	10 *	10 ¶	—
$k_5(s^{-1})$	aa-tRNA accommodation and peptide bond formation	20 *	0.1 *	—
$k_{rej}(s^{-1})$	aa-tRNA release	0.1 *	6 *	—
$k_6 \cdot G^f(s^{-1})$	EF-G binding	4500 §,**	4500 §,**	—
$k_{-6}(s^{-1})$	Reverse EF-G binding	140 §	140 §	—
$k_9(s^{-1})$	Translocation tRNA-mRNA movement	500 §	500 §	—
$k_{11}(s^{-1})$	EF-G and E-site tRNA dissociation	20 §	20 §	—

* Kinetic rate constants from the initial binding until the peptide bound formation obtained from (53) at 20℃.

† Initial binding rate constants are independent of the tRNA-mRNA interaction as it occurs externally to the decoding center (47).

§ Kinetic rate constants for the tRNA-mRNA translocation obtained from (69) at 37℃.

** We use the value of 30 M for the concentration of EF-G as used in previous works (50).

Note: $k_{release}$ Is not present in the table since it is later removed from the model. When used the value is 15s-1 as in (50).

Table C.1.2 aa-tRNA concentrations per binding interaction for the growth rate $0.4h^{-1}$. Individual aa-tRNA concentrations were collected from (59). This table is computed using Table C.2.3 as a guide.

codon	Amino acid	Cognate (c) aa-tRNA concentration [μM]	Near-cognate (nc) aa-tRNA concentration [μM]	Non-cognate (non) aa-tRNA concentration [μM]
aaa	K	6.08	32.17	158.05
aac	N	3.77	35.77	156.76
aag	K	6.08	29.49	160.73
aau	N	3.77	38.66	153.87
aca	T	2.89	35.96	157.45
acc	T	3.78	24.92	167.6
acg	T	4.6	33.42	158.28
acu	T	6.67	36.25	153.38
aga	R	2.74	38.495	155.06
agc	S	4.44	50.77	141.09
agg	R	1.23	31.83	163.24
agu	S	4.44	53.66	138.2
aua	I	5.48	37.21	153.61
auc	I	5.48	29.93	160.89
aug	M	2.23	60.81	133.26
auu	I	5.48	44.94	145.88
caa	Q	2.41	44.69	149.2
cac	H	2.02	43.17	151.11
cag	Q	2.78	48.28	145.24
cau	H	2.02	45	149.28
cca	P	1.83	41.85	152.62
ccc	P	2.27	32.8	161.23
ccg	P	4.67	43.3	148.33
ccu	P	4.1	46.25	145.95
cga	R	15	14.465	166.84
cgc	R	15	32.48	148.82
cgg	R	2.01	49.62	144.67
cgu	R	15	34.31	146.99
cua	L	2.1	57.49	136.71
cuc	L	2.97	48.24	145.09
cug	L	16.21	36.39	143.7
cuu	L	2.97	62.19	131.14

Table C.1.2 continued

codon	Amino acid	Cognate WC aa-tRNA concentration [µM]	Near-cognate (nc) aa-tRNA concentration [µM]	Non-cognate (non) aa-tRNA concentration [µM]
gaa	E	14.88	41.795	139.62
gac	D	7.56	46.78	141.96
gag	E	14.88	45.54	135.88
gau	D	7.56	67.2	121.54
gca	A	10.25	41.135	144.91
gcc	A	1.95	44.02	150.33
gcg	A	10.25	50.15	135.9
gcu	A	10.25	56.65	129.4
gga	G	3.375	72.125	120.8
ggc	G	13.76	44.7	137.84
ggg	G	6.75	57.23	132.32
ggu	G	13.76	65.12	117.42
gua	V	12.12	43.645	140.53
guc	V	3.99	47.11	145.2
gug	V	12.12	63.92	120.26
guu	V	16.11	43.29	136.9
uac	Y	6.41	24.04	165.85
uau	Y	6.41	28.13	161.76
uca	S	4.09	22.04	170.17
ucc	S	2.41	27.87	166.02
ucg	S	5.18	34.52	156.6
ucu	S	6.5	36.8	153
ugc	C	5.01	48.27	143.02
ugg	W	2.98	29.79	163.53
ugu	C	5.01	52.36	138.93
uua	L	3.57	33.1	159.63
uuc	F	3.27	35.88	157.15
uug	L	9.61	41.99	144.7
uuu	F	3.27	52.09	140.94
uaa	stop codon	-	-	-
uag	stop codon	-	-	-
uga	stop codon	-	-	-

C.2 Supplementary Tables for Chapter 2

Table C.2.1 Rate constants for the ribosomal kinetic pathway during translation elongation obtained from experimental sources that discriminate between WC and WB cognate.

Rate constants	Definition	WC	WB	nc	non
$k_1(\mu M^{-1}s^{-1})$	Initial binding	140 *,†	140 *,†	140 *,†	140 *,†
$k_{-1}(s^{-1})$	Reverse initial binding	85 *,†	85 *,†	85 *,†	85 *,†
$k_2(s^{-1})$	Codon reading	190 *	190 *	190 *	—
$k_{-2}(s^{-1})$	Reverse codon reading	0.2 *	1 ‡	80 *	—
$k_3(s^{-1})$	GTPase activation and GTP hydrolysis	260 *	25 ‡	0.4 *	—
$k_4(s^{-1})$	Pi release and EF-Tu rearrangement	10 *	10 ¶	10 ¶	—
$k_5(s^{-1})$	aa-tRNA accommodation and peptide bond formation	20 *	1.6 ‡	0.1 *	—
$k_{rej}(s^{-1})$	aa-tRNA release	0.1 *	1.1 ‡	6 *	—
$k_6 \cdot G^f(s^{-1})$	EF-G binding	4500 §,**	4500 §,**	4500 §,**	—
$k_{-6}(s^{-1})$	Reverse EF-G binding	140 §	140 §	140 §	—
$k_7(s^{-1})$	GTP hydrolysis	100 ‖	100 ‖	100 ‖	—
$k_8(s^{-1})$	Ribosome unlocking	30 ‖	30 ‖	30 ‖	—
$k_9(s^{-1})$	Translocation tRNA-mRNA movement	500 §	500 §	500 §	—
$k_{10}(s^{-1})$	Ribosome re-locking	5 ‖	5 ‖	5 ‖	—
$k_{11}(s^{-1})$	EF-G and E-site tRNA dissociation	20 §	20 §	20 §	—

* Kinetic rate constants from the initial binding until the peptide bound formation obtained from (53) at 20℃.

† Initial binding rate constants are independent of the tRNA-mRNA interaction as it occurs externally to the decoding center (47).

‡ Kinetic rate constants for the cognate WB interaction obtained from (70) at 20℃.

¶ Assumed to be the same as for the cognate WC counterpart due to lack of measurements.

§ Kinetic rate constants for the tRNA-mRNA translocation obtained from (69) at 37℃.

|| Kinetic rate constants for the tRNA-mRNA translocation obtained from (71) at 25℃

** We use the value of 30M for the concentration of EF-G as used in previous works (50).

Table C.2.2 List of wobble pairing nucleotides for each tRNA species derived from the list of codons they recognize as cognates.

tRNA name	tRNA ID	Wobble pairings
Ala1B	1	U, G
Ala2	2	-
Arg2	3	A
Arg3	4	-
Arg4	5	-
Arg5	6	-
Asn	7	U
Asp1	8	U
Cys	9	U
Gln1	10	-
Gln2	11	-
Glu2	12	G
Gly1	13	-
Gly2	14	G
Gly3	15	U
His	16	U
Ile1	17	U
Ile2	18	A
Leu1	19	-
Leu2	20	U
Leu3	21	G
Leu4	22	-
Leu5	23	G
Lys	24	G
Metm	25	-
Phe	26	U
Pro1	27	-
Pro2	28	U
Pro3	29	U, G
Ser1	30	U, G
Ser2	31	-
Ser3	32	U
Ser5	33	U
Thr1	34	U
Thr2	35	-
Thr3	36	U
Thr4	37	U, G
Trp	38	-
Tyr1	39	U
Tyr2	40	U
Val1	41	G, U
Val2A	42	U
Val2B	43	U

Table C.2.3 List of tRNA species that are cognate WC, cognate WB and near-cognate to each codon. The remaining tRNA species not listed for each codon are non-cognate. This list is used to compute tRNA concentrations for each codon per type of binding interaction. We do not use information on tRNA concentrations for the three stop codons uaa, uag and uga because the termination rates are fixed for each gene in the simulations and do not depend on the availability of translational resources.

codon (id)	Cognate WC	Cognate WB	Near-cognate (nc)
aaa (1)	24	-	5, 7, 10, 12, 18, 37
aac (2)	7	-	8, 16, 17, 24, 32, 34, 36, 39, 40
aag (3)	-	24	6, 7, 11, 12, 25, 35, 37
aau (4)	-	7	8, 16, 17, 24, 32, 34, 36, 37, 39, 40
aca (5)	37	-	1, 5, 18, 24, 29, 30, 34, 35, 36
acc (6)	34, 36	-	2, 7, 17, 28, 32, 33, 35, 37
acg (7)	35	37	1, 6, 24, 25, 27, 29, 30, 31, 34, 36
acu (8)	-	34, 36, 37	1, 7, 17, 28, 29, 30, 32, 33, 35
aga (9)	5	-	3, 6, 14, 18, 24, 32, 32, 37
agc (10)	32	-	3, 5, 6, 7, 9, 15, 17, 34, 36
agg (11)	6	-	4, 5, 13, 14, 24, 25, 32, 35, 37, 38
agu (12)	-	32	3, 5, 6, 7, 9, 15, 17, 34, 36, 37
aua (13)	-	18	5, 17, 21, 23, 24, 25, 37, 41
auc (14)	17	-	7, 18 20, 25, 26, 32, 34, 36, 42, 43
aug (15)	25	-	6, 17, 18, 19, 21, 22, 23, 24, 35, 37, 41
auu (16)	-	17	7, 18, 20, 25, 26, 32, 34, 36, 37 41, 42, 43
caa (17)	10	-	3, 11, 12, 16, 21, 24, 29
cac (18)	16	-	3, 7, 8, 10, 11, 20, 28, 39, 40
cag (19)	11	-	4, 10, 12, 16, 19, 21, 24, 27, 29
cau (20)	-	16	3, 7, 8, 10, 11, 20, 28, 29, 39, 40
cca (21)	29	-	1, 3, 10, 21, 27, 28, 30, 37
ccc (22)	28	-	2, 3, 16, 20, 27, 29, 33, 34, 36
ccg (23)	27	29	1, 4 11 19 21 28 30 31 35 37
ccu (24)	-	28, 29	1, 3, 16, 20, 27, 30, 33, 34, 36, 37
cga (25)	-	3	4, 5, 10, 14, 21, 29, 32
cgc (26)	3	-	4, 9, 15, 16, 20, 28, 32
cgg (27)	4	-	3, 6, 11, 13, 14, 19, 21, 27, 29, 38
cgu (28)	3	-	4, 9, 15, 16, 20, 28, 29, 32
cua (29)	21	-	3, 10, 18, 19, 20, 23, 29, 41
cuc (30)	20	-	3, 16, 17, 19, 21, 26, 28, 42, 43
cug (31)	19	21	4, 11, 20, 22, 23, 25, 27, 29, 41
cuu (32)	-	20	3, 16, 17, 19, 21, 26, 28, 29, 41, 42, 43
gaa (33)	12	-	1, 8, 10, 14, 24, 41
gac (34)	8	-	2, 7, 12, 15, 16, 39, 40, 42, 43
gag (35)	-	12	1, 8, 11, 13, 14, 24, 41

Table C.2.3 continued

codon (id)	Cognate WC	Cognate WB	Near-cognate (nc)
gau (36)	-	8	1, 7, 12, 15, 16, 39, 40, 41, 42, 43
gca (37)	1	-	2, 12, 14, 29, 30, 37, 41
gcc (38)	2	-	1, 8, 15, 28, 33, 34, 36, 42, 43
gcg (39)	-	1	2, 12, 13, 14, 27, 29, 30, 31, 35, 37, 41
gcu (40)	-	1	2, 8, 15, 28, 29, 30, 33, 34, 36, 37, 41, 42, 43
gga (41)	14	-	1, 3, 5, 12, 13, 15, 32, 41
ggc (42)	15	-	2, 3, 8, 9, 13, 14, 32, 42, 43
ggg (43)	13	14	1, 4, 6, 12, 15, 38, 41
ggu (44)	-	15	1, 3, 8, 9, 13, 14, 32, 41, 42, 43
gua (45)	41	-	1, 12, 14, 18, 21, 23, 42, 43
guc (46)	42, 43	-	2, 8, 15, 17, 20, 26, 41
gug (47)	-	41	1, 12, 13, 14, 19, 21, 22, 23, 25, 42, 43
guu (48)	-	41, 42, 43	1, 8, 15, 17, 20, 26
uac (49)	39, 40	-	7, 8, 9, 16, 26, 33
uau (50)	-	39, 40	7, 8, 9, 16, 26, 30, 33
uca (51)	30	-	1,23, 29, 32, 31, 33, 37
ucc (52)	33	-	2, 9, 26, 28, 30, 31, 34, 36, 39, 40
ucg (53)	31	30	1, 22, 23, 27, 29, 33, 35, 37, 38
ucu (54)	-	30, 33	1, 9, 26, 28, 29, 31, 34, 36, 37, 39, 40
ugc (55)	9	-	3, 15, 26, 32, 32, 33, 38, 39, 40
ugg (56)	38	-	4, 6, 9, 13, 14, 22, 23, 32, 30, 31
ugu (57)	-	9	3, 15, 26, 32, 30, 32, 33, 38, 39, 40
uua (58)	23	-	18, 21, 22, 26, 32, 30, 41
uuc (59)	26	-	9, 17, 20, 22, 23, 33, 39, 40, 42, 43
uug (60)	22	23	19, 21, 25, 26, 30, 31, 38, 41
uuu (61)	-	26	9, 17, 20, 22, 23, 30, 33, 39, 40, 41, 42, 43

Table C.2.4 aa-tRNA concentrations per binding interaction estimated for the growth rate 1.07h^{-1}. Individual aa-tRNA concentrations were collected from (59) and estimated as explained in methods.

codon (id)	Amino acid	Cognate WC aa-tRNA concentration [µM]	Cognate WB aa-tRNA concentration [µM]	Near-cognate (nc) aa-tRNA concentration [µM]	Non-cognate (non) aa-tRNA concentration [µM]
aaa	K	5.71	0.00	29.52	151.82
aac	N	3.09	0.00	33.19	150.77
aag	K	0.00	5.71	28.00	153.34
aau	N	0.00	3.09	35.79	148.17
aca	T	2.59	0.00	34.88	149.57
acc	T	3.39	0.00	23.86	159.80
acg	T	1.83	2.59	34.15	148.48
acu	T	0.00	5.98	36.98	144.09
aga	R	2.48	0.00	34.28	150.28
agc	S	3.70	0.00	49.29	134.07
agg	R	1.51	0.00	29.56	155.98
agu	S	0.00	3.70	51.88	131.47
aua	I	0.00	4.08	34.12	148.85
auc	I	6.67	0.00	26.79	153.59
aug	M	2.19	0.00	57.75	127.12
auu	I	0.00	6.67	38.20	142.18
caa	Q	3.09	0.00	41.57	142.40
cac	H	1.92	0.00	38.96	146.17
cag	Q	2.60	0.00	47.49	136.96
cau	H	0.00	1.92	40.59	144.54
cca	P	1.63	0.00	41.75	143.67
ccc	P	1.89	0.00	30.41	154.75
ccg	P	2.90	1.63	43.79	138.73
ccu	P	0.00	3.52	45.45	138.07
cga	R	0.00	13.29	14.12	159.64
cgc	R	13.29	0.00	30.84	142.92
cgg	R	1.28	0.00	48.48	137.29
cgu	R	13.29	0.00	32.47	141.29
cua	L	2.23	0.00	52.07	132.75
cuc	L	3.20	0.00	47.69	136.16
cug	L	14.54	2.23	32.38	137.90
cuu	L	0.00	3.20	58.14	125.70
gaa	E	14.19	0.00	38.61	134.25
gac	D	6.47	0.00	44.52	136.06
gag	E	0.00	14.19	41.65	131.21
gau	D	0.00	6.47	62.78	117.80
gca	A	11.12	0.00	36.95	138.97
gcc	A	1.68	0.00	43.58	141.79
gcg	A	0.00	11.12	46.27	129.66
gcu	A	0.00	11.12	51.82	124.11
gga	G	3.40	0.00	67.61	116.03
ggc	G	14.19	0.00	40.85	132.01
ggg	G	3.53	3.40	53.94	126.17
ggu	G	0.00	14.19	59.11	113.75
gua	V	8.82	0.00	42.56	135.67
guc	V	4.12	0.00	44.06	138.87
gug	V	0.00	8.82	65.08	113.15
guu	V	0.00	12.94	44.68	129.43

Table C.2.4 continued

codon (id)	Amino acid	Cognate WC aa-tRNA concentration [μM]	Cognate WB aa-tRNA concentration [μM]	Near-cognate (nc) aa-tRNA concentration [μM]	Non-cognate (non) aa-tRNA concentration [μM]
uac	Y	5.34	0.00	21.57	160.14
uau	Y	0.00	5.34	26.21	155.51
uca	S	4.64	0.00	22.23	160.18
ucc	S	2.40	0.00	25.68	158.97
ucg	S	1.06	4.64	35.08	146.28
ucu	S	0.00	7.04	34.71	145.30
ugc	C	4.67	0.00	44.77	137.61
ugg	W	2.84	0.00	29.84	154.37
ugu	C	0.00	4.67	49.41	132.98
uua	L	3.42	0.00	29.12	154.51
uuc	F	3.02	0.00	36.15	147.87
uug	L	6.33	3.42	39.34	137.96
uuu	F	0.00	3.02	49.61	134.42
uaa	stop codon	-	-	-	-
uag	stop codon	-	-	-	-
uga	stop codon	-	-	-	-

Table C.2.5 Parameter values used for computing cell volume and average number of mRNA copies per *E. coli* cell at four different growth rates.

Growth Rate [h^{-1}]	**0.4**	**0.7**	**1.07**	**1.6**
Total ribosomes per cell (R^T) †	5705	8197	14456	29551
Total tRNA per cell ($tRNA^T$) †	62130*	75868	133925	274138
mRNA synthesis rate per cell † [Nucleotides/min/cell]	274960	572530	939533	1465240
τ_{mRNA} [min]	1	1	1	1
nt_{mRNA}	951	951	951	951
Cell volume (V_{cell}) [10^{-16} ·L] ‡	5.26	6.04	9.82	15.1
M^T [mRNA molecules/cell] ‡	289	602	988	1541

* Total tRNA per cell exceptionally obtained from the sum of the number of molecules of each tRNA species measured and reported directly in (59).
† Values estimated for the growth rates of interest from the fitting of the data obtained in (58).
‡ Values computed from Eq. A.2.1 and Eq. A.2.2.
τ_{mRNA}: Average functional life of mRNA.
nt_{mRNA}: Average number of nucleotides for the mRNA sequences in *E. coli*.

Table C.2.6 Comparison between mRNA sequencing data from *E. coli* at low and high growth rates obtained from (79, 80). The mRNA species expressed at each growth rate were separated into two groups: (i) commonly expressed mRNA species and (ii) uniquely expressed mRNA species. Statistics from the mRNA copy number distributions at each growth rate condition and per group are shown. We also compare the mRNA codon usage frequency (mCU) at each growth rate with the CU from *E. coli K12*, as well as the respective average mRNA lengths. The copy number levels at high growth rate were normalized by the estimated amount of total mRNA copies for the respective growth rate, computed from data in (58). The copy number levels from the data at low growth rate had been calibrated by the single cell measurements performed in the same paper and normalization with respect to the total number of mRNA copies in the cell was not necessary. In order to compute mCU for each growth rate we have rounded up the mRNA copy levels to the nearest integer such that values smaller than 0.5 would be represented by one copy. For the analysis, we excluded the mRNAs that had zero expression levels.

	Distribution features for the same mRNAs (i)		**Distribution features for the uniquely expressed mRNAs (ii)**	
Growth Rate [h^{-1}]	0.4*	>2.5†	0.4*	>2.5†
Number of mRNA species expressed	603	603	224	1784
Maximal copy number value	6.5789	28.6732	6.0843	96.9363
Minimal copy number value	0.0011	0.0072	3.8880e-04	0.0048
Mean of mRNA copy number distribution	0.2205	0.7411	0.0986	1.1284
Median of copy number distribution	0.0727	0.2260	0.0175	0.1430
Standard deviation of copy number distribution	0.5575	2.0105	0.4352	4.7472

Comparison with *E. coli K12*

	Mean gene length [codons]	**Mean value of individual relative deviations between mCU and CU from *E. coli K12***
E. coli K12	317	-
0.4 h^{-1}*	338	15% (±9.73)
>2.5 h^{-1}†	319	23% (±13.37)

* mRNA sequencing data from (79).
† mRNA sequencing data from (80).

Table C.2.7 Detailed mRNA Luciferase transcripts used for recording ribosome occupancy time lags and in the study of synonymous sequence optimization. All Luciferase sequences code for the same amino acid sequence and apart from the changes in synonymous codons they all have the same C-terminal c-myc-His6 epitope tag as in (60).

Type of study	Protein	mRNA sequence
Wild Type (Sequence from (60)) also used to compute ribosome occupancy time lags	Firefly Luciferase	auggaagacgccaaaaacauaaagaaaggcccggcgccauucuauccucuagaggauggaaccgcugg agagcaacugcauaaggcuaugaagagauacgcccugguuccuggaacaauugcuuuuacagaugcac auaucgaggugaacaucacguacgcggaauacuucgaaauguccguucgguuggcagaagcuaugaaa cgauaugggcugaauacaaaucacagaaucgucguaugcagugaaaacucucuucaaauccuuuaugcc ggguguugggcgcguuauuuuaucggaguugcaguugcgcccgcgaacgacauuuauaaaugaacgugaa uugcucaacaguaugaacauuucgcagccuaccguaguguuuguuuccaaaaagggguugcaaaaaau uuugaacgugcaaaaaaauuaccaauaauccagaaaauuauuaucauggauucuaaaacggauuacc agggauuucagucgauguacacguucgucacaucucaucuaccucccgguuuuaaugaauacgauuu uguaccagaguccuuugaucgugacaaaacaauugcacugauaaugaauuccucuggaucuacuggg uuaccuaagggugugggcccuuccgcauagaacugccugcgucagauucucgcaugccagagauccuau uuuuggcaaucaaaucauuccggauacugcgauuuuaaguguuguuccauuccaucacgguuuugga auguuuacuacacucggauauuugauauguggauuucgagucgucuuaauguauagauuugaagaag agcuguuuuuacgaucccuucaggauuacaaaauucaaagugcguugcuaguaccaacccuauuuuca uucuucgccaaaagcacucugauugacaaauacgauuuaucuaauuuacacgaaauugcuucuggggg cgcaccucuuucgaaagaagucggggaagcgguugcaaaacgcuuccaucuuccagggauacgacaag gauaugggcucacugagacuacaucagcuauucugauuacacccgaggggugauaaaccgggcgcg gucgguaaagguuguuccauuuuuugaagcgaagguuguggaucuggauaccgggaaaacgcugggcg uuaaucagagaggcgaauuaugugucagaggaccuaugauuauguccgguuauguaaacaauccgga agcgaccaacgccuugauugacaaggauggauggcuacauucuggagacauagcuuacugggacgaag acgaacacuucuucauaguugaccgcuugaagucuuuaauuaaauacaaaggauaucaggugccccc gcugaauuggaaucgauauuguuacaacaccccaacaucuucgacgcgggcguggcaggucuucccga cgaugacgccggugaacuucccgccgccguuguuguuuuggagcacggaaagacgaugacggaaaaa gagaucguggauuacgucgccagucaaguaacaaccgcgaaaaaguugcgcggaggaguuguguuug uggacgaaguaccgaaaggucuuaccggaaaacucgacgcaagaaaaaucagaggaguccucauaaag gccaagaaagggcggaaaguccaaauugaucgaaggccgcggaucugguacuaguggcgggucaggug gcucggggcgauccgaacaaaaacuuauuucugaagaagacuugcaccaucaccaucaccauuaa
WC & tRNA genes (Sequence from (60))		auggaagacgcaaaaaacaucaaaaaaggcccggcaccguucuacccgcuggaagacggcaccgcaggc gaacagcugcacaaagcaaugaaacguuacgcacugguaccgggcaccaucgcauucaccgacgcacac aucgaaguaaacaucaccuacgcagaauacuucgaaauguccguacgucuggcagaagcaaugaaacg uuacggccugaacaccaaccaccguaucguaguaugcuccgaaaacucccugcaguucuucaugccgg uacugggcgcacuguucaucggcguagcaguagcaccggcaaacgacaucuacaacgaacgugaacug cugaacuccaugaacaucucccagccgaccguaguauucguauccaaaaaaggccugcagaaaauccug aacguacagaaaaaacugccgaucauccagaaaaucaucaucauggacuccaaaaccgacuaccagggc uuccaguccauguacaccuucguaacccucccaccugccgccgggcuucaacgaauacgacuucguacc ggaauccuucgaccgugacaaaaccaucgcacugaucaugaacuccuccggcuccaccggccugccgaa aggcguagcacugccgcaccguaccgcaugcguacguuucucccacgcacgugacccgaucuucggca accagaucaucccggacaccgcaauccuguccguaguaccguuccaccacggcuucggcauguucacc acccugggcuaccugaucugcggcuuccguguaguacugauguaccguuucgaagaagaacuguucc ugcguucccugcaggacuacaaaauccagucgcacugcugguaccgacccuguucuccuucuucgca aaauccacccugaucgacaaauacgaccuguccaaccugcacgaaaucgcauccggcggcgcaccgcug uccaaagaaguaggcgaagcaguagcaaaacguuuccaccugccgggcauccgucagggcuacggccu gaccgaaaccaccuccgcaauccugaucaccccggaaggcgacgacaaaccgggcgcaguaggcaaagu aguaccguucuucgaagcaaaaguaguagaccuggacaccggcaaaacccugggcguaaaccagcgug gcgaacugugcguacguggcccgaugaucauguccggcuacguaaacaacccggaagcaaccaacgca cugaucgacaaagacggcuggcugcacuccggcgacaucgcauacugggacgaagacgaacacuucuu caucguagaccgucugaaaucccugaucaaauacaaaggcuaccagguagcaccggcagaacuggaau ccauccugcugcagcacccgaacaucuucgacgcaggcguagcaggccugccggacgacgacgcaggc gaacugccggcagcaguaguaguacuggaacacggcaaaaccaugaccgaaaaagaaaucguagacuac guagcaucccagguaaccaccgcaaaaaaacugcguggcggcguaguauucguagacgaaguaccgaa aggccugaccggcaaacuggacgcacguaaaaauccgugaaauccugaucaaagcaaaaaaaggcggcaa auccaaacugaucgaaggccgcggaucugguacuaguggcgggucagguggcucggggcgauccgaa caaaaacuuauuucugaagaagacuugcaccaucaccaucaccauuaa
CU based (Sequence from (60))		auggaagaugcgaaaaacauuaaaaaaggcccggcgccguuuuaucgcuggaagauggcaccgcgggg cgaacagcugcauaaagcgaugaaacgcuaugcgcuggugccgggcaccauugcguuuaccgaugcgc auauugaagugaacauuaccuaugcggaauauuuugaaaugagcgugcgccuggcggaagcgaugaa acgcuauggccugaacaccaaccaucgcauuguggugugcagcgaaaacagccugcaguuuuuuaugc

		cggugcugggcgcgcuguuuauuggcguggcgguggcgccggcgaacgauauuuauaacgaacgcga acugcugaacagcaugaacauuagccagccgaccgugguguuugugagcaaaaaaggccugcagaaaa uucugaacgugcagaaaaaacugccgauuauucagaaaauuauuauuauuggauagcaaaaccgauuau cagggcuuucagagcauguauaccuuugugaccagccaucugccgccgggcuuuaacgaauaugauu uugugccggaaagcuuugaucgcgauaaaaccauugcgcugauuaugaacagcagcggcagcaccggc cugccgaaaggcguggcgcugccgcaucgcaccgcgugcgugcgcuuuagccaugcgcgcgauccga uuuuuggcaaccagauuauuccggauaccgcgauucugagcgugguguccguuucaucauggcuuugg cauguuuaccacccuggguauucugauuugcggcuuucgcguggugcugauguaucgcuuugaagaa gaacuguuucugcgcagccugcaggauuauauaaaauucagagcgcgcugcuggugccgacccuguuua gcuuuuuugcgaaaagcacccugauugauaaaauaugaucugagcaaccugcaugaaauugcgagcggc ggcgcgccgcugagcaaagaaguggcgaagcgguggcgaaacgcuuucaucugccgggcauucgcc agggcuauggccugaccgaaaccaccagcgcgauucugauuaccccggaaggcgaugauaaaccgggc gcgguggcaaaguggugccguuuuuugaagcgaaaguggugguaucuggauaccggcaaaacccugg gcgugaaccagcgcggcgaacugugcgugcgcggcccgaugauuaugagcggcuaugugaacaaccc ggaagcgaccaacgcgcugauugauaaagauggcuggcugcauagcggcgauauugcguauuggga gaagaugaacauuuuuuuauuguggaucgccugaaaagccugauuaaauauaaaggcuaucaggugg cgccggcggaacuggaaagcauucugcugcagcauccgaacauuuuugaugcgggcguggcgggccu gccggaugaugaugcgggcgaacugccggcggcgguggguggugcuggaacauggcaaaaccaugacc gaaaaagaaauuguggauuauguggcgagccaggugaccaccgcgaaaaaacugcgcggcggcgugg uguuuguggaugaagugccgaaaggccugaccggcaaacuggaugcgcgcaaaauucgcgaaauucu gauuaaagcgaaaaaaggcggcaaaagcaaacugaucgaaggccgcggaucugguacuaguggcgggu cagguggcucggggcgauccgaacaaaaacuuauuucugaagaagacuugcaccaucaccaucaccau uaa
WB based (Sequence from (60))		auggaggaugcuaagaauauaaagaaggguccugcuccuuuuuauccucuugaggaugguacugcug gugagcaacuucauaaggcuaugaagcgauaugcucuuguuccugguacuauagcuuuuacugaugc ucauauagagguuaauauaacuuaugcugaguauuuugagaugucuguucgacuugcugaggcuaug aagcgauaugucuuaauacuaaucaucgaauaguuguuuguucugagaauucucuucaauuuuuua ugccuguucuuggugcucuuuuuauagguguugcugcuucgcuaaugauauauauaaaugagc gagagcuucuuaaauucuaugaauauaucucaaccuacugcuuguuuuuguuucuaagaagggcuuuca aaagauacuuaaauguucaaaagaagcuuccuauaauacaaaagauaauaauaaauggauucuaagacug auuaucaaggcuuuucaaucuauguauacuuuuguuacuucucaucuuccuccuggcuuuuaaugagua ugauuuuguuccugagucuuuugaucgagauaagacuauagcucuuauaaaugaauucuucuggcuucu acuggucuuccuaagggcuguugcucuuccucaucgaacugcuuguguucgauuuucucaugcucgag auccuauauuuggcuaaucaaauaauaccugauacugcuauacuuucuguuguuccuuuucaucaugg uuuuggcuauguuuacuacucuuggcuuaucuuauaugcuggcuuuucgaguuguucuuauguaucgauu ugaggaggagcuuuuucuucgaucucuucaagauuauaagauacaaucugcucuucuuguuccuacu cuuuuuucuuuuuuugcuaagucuacucuuauagauaaguaugauccuuucuaaucuucaugagauag cuucuggcuggugcuccucuuucuaaggagguuggugaggcuguugcuaagcgauuucaucuuccug guauacgacaaggcuuaugcuucuacugagacuacuucugcuauacuuauaaacuccugagggugauga uaagccuggugcuguuggcuaaggcuuguuccuuuuuuugaggcuaaggcuuguugaucuugauacugg uaagacucuuggcuguuaaucaacgaggugagcuuuguguucgaggcccuaugauaaugucuggcuua uguuaauaaaccugaggcuacuaaugcucuuauagauaaggauggcuuggcuucauucuggugauaua gcuuauuggcaugaggaugagcauuuuuuuuauaguugaucgacuuaagucucuuauaaaguauaaagg guuaucaaguugcuccugcugagcuugagucuauacuucuucaacauccuaauauauuugaugcugg uguugcuggucuuccugaugaugaugcuggugagcuuccugcugcuguuguuguucuugagcaugg uaagacuaugacugagaaggagauaguuugauuauguugcuucucaaguuacuacugcuaagaagcuu cgaggugguguuguuuuuguugaugaggcuuccuaagggucuuacuggcuaagcuugaugcucgaaag auacgagagauacuuauaaaggcuaagaaggguggcuaagucuuaucgaaggccgcggaucug guacuaguggcgggucaggcuggcucggggcgauccgaacaaaaacuuauuucugaagaagacuugca ccaucaccaucaccauuaa
k_{eff}^{max} **based**		auggaagacgcaaaaaacaucaaaaaaggcccggcaccguucuacccgcuggaagacggcaccgcaggc gaacaacugcacaaagcaaugaaacgcuacgcacuggcuaccgggcaccaucgcauucaccgacgcacac aucgaaguaaacaucaccuacgcagaauacuucgaaaugucaguacgccuggcagaagcaaugaaacgc uacggccugaacaccaaccaccgcaucguaguaugcucagaaaacucacugcaauucuucaugccggu acugggcgcacuguucaucggcguagcaguagcaccggcaaacgacaucuacaacgaacgcgaacugc ugaacucaaugaacaucucacaaccgaccguaguauucguaucaaaaaaaggccugcaaaaaauccuga acguacaaaaaaaacugccgaucauccaaaaaaucaucaucauggacucaaaaaccgacuaccaaggcu uccaaucaauguacaccuucguaaccucacaccugccgccgggcuucaacgaauacgacuucguaccg gaaucauucgaccgcgacaaaaccaucgcacugaucaugaacucaucaggcucaaccggccugccgaaa ggcguagcacugccgcaccgcaccgcaugcguacgcuucucacacgcacgcgacccgaucuucggcaa ccaaaucaucccggacaccgcaauccugucaguaguaccguuccaccacggcuucggcauguucaccac ccugggcuaccugaucugcggcuuccgcguaguacugauguaccgcuucgaagaagaacuguuccug cgcucacugcaagacuacaaaauccaaucagcacugcugguaccgacccuguucucauucuucgcaaaa ucaacccugaucgacaaauacgaccugucaaaccugcacgaaaucgcaucaggcggcgcaccgcugcua aaagaaguaggcgaagcaguagcaaaacgcuuccaccugccgggcauccgccaaggcuacggccugac

		cgaaaccaccucagcaauccugaucaccccggaaggcgacgacaaaccgggcgcaguaggcaaaguagu accguucuucgaagcaaaaguaguagaccuggacaccggcaaaacccugggcguaaaccaacgcggcg aacugugcguacgcggcccgaugaucaugucaggcuacguaaacaacccggaagcaaccaacgcacug aucgacaaagacggcuggcugcacucaggcgacaucgcauacugggacgaagacgaacacuucuucau cguagaccgccugaaaucacugaucaaauacaaaggcuaccaaguagcaccggcagaacuggaaucaau ccugcugcaacacccgaacaucuucgacgcaggcguagcaggccugccggacgacgacgcaggcgaac ugccggcagcaguaguaguacuggaacacggcaaaaccaugaccgaaaaagaaaucguagacuacgua gcaucacaaguaaccaccgcaaaaaaacugcgcggcggcguaguauucguagacgaaguaccgaaagg ccugaccggcaaacuggacgcacgcaaaauccgcgaaauccugaucaaagcaaaaaaaggcggcaaauc aaaacugaucgaaggccgcggaucugguacuaguggcgggucagguggcucggggcgauccgaacaa aaacuuauuucugaagaagacuugcaccaucaccaucaccauuaa
WC based		auggaagacgccaaaaacaucaaaaaaggaccagccccauucuacccacucgaagacggaacagccgga gaacagcuccacaaagccaugaaacgcuacgcccucgucccaggaacaaucgccuucacagacgcccac aucgaagucaacaucacauacgccgaauacuucgaaaugagcguccgccucgccgaagccaugaaacgc uacggacucaacacaaaccaccgcaucgucgucugcagcgaaaacagccuccaguucuucaugccaguc cucggagcccucuucaucggagucgccgucgccccagccaacgacaucuacaacgaacgugaacuccuc aacagcaugaacaucagccagccaacagucgucuucgucagcaaaaaaggacuccagaaaauccucaac guccagaaaaaacucccaaucaucccagaaaaucaucaucauggacagcaaaacagacuaccagggauuc cagagcauguacacauucgucacaagccaccucccaccaggauucaacgaauacgacuucgucccagaa agcuucgaccgugacaaaacaaucgcccucaucaugaacagcagcggaagcacaggacucccaaaagga gucgcccucccacaccgcacagccugcguccgcuucagccacgcccgcgacccaaucuucggaaaccag aucaucccagacacagccauccucagcgucgucccauuccaccacggauucggaauguucacaacacuc ggauaccucaucugcggauuccgcgucguccucauguaccgcuucgaagaagaacucuuccuccgcag ccuccaggacuacaaaauccagagcgcccuccucgucccaacacucuucagcuucuucgccaaaagcac acucaucgacaaaauacgaccucagcaaccuccacgaaaucgccagcggaggagccccacucagcaaaga agucggagaagccgucgccaaacgcuuccaccucccaggaauccgccagggauacggacucacagaaac aacaagcgccauccucaucacaccagaaggagacgacaaaccaggagccgucggaaaagucgucccauu cuucgaagccaaagucgucgaccucgacacaggaaaaacacucggagucaaccagcgcggagaacucug cguccgcggaccaaugaucaugagcggauacgucaacaacccagaagccacaaacgcccucaucgacaa agacggauggcuccacagcggagacaucgccuacugggacgaagacgaacacuucuucaucgucgacc gccucaaaagccucaucaaauacaaaggauaccaggucgccccagccgaacucgaaagcauccuccucc agcacccaaacaucuucgacgccggagucgccggacucccagacgacgacgccggagaacucccagccg ccgucgucguccucgaacacggaaaaacaaugacagaaaaagaaaucgucgacuacgucgccagccagg ucacaacagccaaaaaacuccgcggaggagucgucuucgucgacgaagucccaaaaggacucacaggaa aacucgacgcccgcaaaauccgcgaaauccucaucaaagccaaaaaaggaggaaaaagcaaacucaucg aaggccgcggaucugguacuaguggcgggucagguggcucggggcgauccgaacaaaaacuuauuuc ugaagaagacuugcaccaucaccaucaccauuaa
TC based		

C.3 Supplementary Tables for Chapter 3

Table C.3.1 Summary of Web services provided per database and possible query results of interest. HMDB and SEED do not have web services but are included here to show what information we can retrieve from them.

		Search inputs			*Outputs from web service request*			
Database	**Web services**	Search by compound name	Search by DB identifier	Search by external identifiers	Compound structure information (InChI, SMILE, molfile, etc)	Compound synonyms	Compound formula	Cross-reference to external identifiers
KEGG	X	X	X	-	X	X	X	X
ChEBI	X	X	X	-	X	X	X	X
LipidMaps	X	X	X	-	X	X	X	X
Pubchem	X	X	X	X	X	X	X	X
BiGG	X	X	X	-	-	X	X	X
MetaCyc	X	-	X	X	X	X	X	X
HMDB	-	X	X	X	X	X	X	X
SEED	-	X	X	X	X	X	X	X

Table C.3.2 Reactions that present $\Delta_r G'^{\circ}$ sign switched due to a change in ionic strength value from 0.15 M to 0.25 M. From the 3671 reactions in Recon 2 v4 with thermodynamic constrains successfully added. Despite the change in sign and value of $\Delta_r G'^{\circ}$, the uncertainty associated with the $\Delta_r G'^{\circ}$ estimation is sufficiently high such that the bounds for $\Delta_r G'$estimated for a range of physiological metabolite concentrations are approximately the same and the reactions are reversible. These reactions are not transports and take place in the mitochondria (4) and 1 in the endoplasmic reticulum.

RXN identifier	**ΔrG'° (I=0.15)**	**ΔrG'° (I=0.25)**	**Reaction name**	**Reaction formula**
MMCDm	-0.0489	0.2208	C-3 sterol dehydrogenase (4-methylzymosterol)	proton + (S)-methylmalonyl-CoA(5-) <=> carbon dioxide + Propanoyl-CoA
r1137	0.0068	-0.0470	Methylmalonyl-CoA decarboxylase, mitochondrial	Nicotinamide adenine dinucleotide + 4alpha-Methylzymosterol-4-carboxylate <=> carbon dioxide + Nicotinamide adenine dinucleotide - reduced + 3-Keto-4-methylzymosterol
C3STDH1r	0.0068	-0.0470	nucleoside-diphosphate kinase (ATP:dIDP), mitochondrial	4-Methylzymosterol intermediate 1 + Nicotinamide adenine dinucleotide <=> 4-Methylzymosterol intermediate 2 + carbon dioxide + Nicotinamide adenine dinucleotide - reduced
NDPK10m	0.0298	-0.0363	nucleoside-diphosphate kinase (ATP:IDP), mitochondrial	ATP + dIDP <=> ADP + dITP(4-)
NDPK9m	0.0298	-0.0363	NAD(P) dependent steroid dehydrogenase-like EC:1.1.1.170	ATP + IDP(3-) <=> ADP + ITP(3-)

Table C.3.3 Listing of compounds that did not have common external database identifiers between GEMap using compound name search and MetaNetX (MNX) results. Problematic mismatches for the purpose of thermodynamics computations are highlighted in orange.

Comments	Compound name in Recon 2 v4	MNX results	GEMap results
2,4-Dihydroxy-nitrophenol compound in MNX. Structure is different by 1 H, proton and double bound.	2-hydroxy-4-nitrophenolate	HMDB06200|24nph	CHEBI:57730|HMDB62694|6971250
CHEBI:27718 in ChEBI DB is linked to G00058	Type IIIH glycolipid	CHEBI:27718|CHEBI:22090|CHEBI:9797|fucgalacgalfuc12gal14acglcgalgluside_hs	G00058|cpd21546|fucgalacgalfuc12gal14acglcgalgluside_cho
CHEBI:28574 in ChEBI DB is linked to G00059	Type IIIA glycolipid	CHEBI:28574|CHEBI:22089|CHEBI:9796|acgalfucgalacgalfuc12gal14acglcgalgluside_hs	G00059|cpd21547
3a,7a,12a-Trihydroxy-5b-cholestanoyl-CoA in MNX. Recon 2 v4 name is synonym of this name. Differ in stereochemistry layer and protonation	(25R)-3alpha,7alpha,12alpha-trihydroxy-5beta-cholestan-26-oyl-CoA(4-)	C04760|CHEBI:15493|CHEBI:11899|CHEBI:1699|CHEBI:20222|CHEBI:63001|HMDB02178|cholcoar|LMST01010218|cpd02898	CHEBI:58677|45266721
CHEBI:63085 in ChEBI DB is linked to G00069	nLc8Cer	CHEBI:63085|galacglc13galacglcgal14acglcgalgluside_hs	G00069|cpd21556|galacglc13galacglcgal14acglcgalgluside_cho
MNX CHEBI identifiers are no longer available in ChEBI	Lea glycolipid	CHEBI:28246|CHEBI:21431|CHEBI:6396|fuc14galacglcgalgluside_hs	G00046|cpd21537|fuc14galacglcgalgluside_cho
CHEBI:62562 in ChEBI DB is linked to G00056	Ley glycolipid	CHEBI:62562|fucfuc12gal14acglcgalgluside_hs	G00056|cpd21544|fucfuc12gal14acglcgalgluside_cho
Umbelliferone as namesd in the model is a hydroxycoumarin (compound identified in MNX) substituted by a hydroxyl group at position 7	umbelliferone	CHEBI:37912|CHEBI:24691|CHEBI:24692|hcoumarin	CHEBI:27510|C09315|HMDB29865|5281426|cpd06210
Epiandrosterone compound in MNX	16alpha-hydroxydehydroepiandrosterone	C07635|CHEBI:541975|CHEBI:4802|HMDB00365|eandrstrn|LMST02020023|cpd04797	CHEBI:27771|C05139|LMST02020064|102030|HMDB00352|cpd03059
20-hydroxycholesterol name is repeated in Recon 2 for two different compounds and reactions. 7a-Hydroxycholesterol in MNX as it should be in Recon 2 v4 model for one of the metabolites. However, the deltaG is similar for both	20-hydroxycholesterol	C03594|CHEBI:17500|CHEBI:12263|CHEBI:13980|CHEBI:20801|CHEBI:2293|CHEBI:58167|CHEBI:42989|CHEBI:35351|CHEBI:42983|HMDB01496|HMDB06119|HMDB59592|xol7a|LMST01010013|LMST01010047|cpd02262	CHEBI:1296|HMDB06283|C05500|440711|LMST01010201|121935|cpd03274
9-cis-retinal in MNX differs in stereochemistry layer	all-trans-retinal	C16681|CHEBI:78273|HMDB06218|retinal_cis_9|LMPR01090017|cpd16479	CHEBI:17898|HMDB01358|C00376|638015|LMPR01090002|cpd00304
1-lauroyl-*sn*-glycerol 3-phosphate in MNX, whereas 1-acyl-sn-glycerol 3-phosphate(2-) identified in ChEBI has R groups	1-acyl-sn-glycerol 3-phosphate(2-)	CHEBI:62840|CHEBI:72682|1ddecg3p|LMGP10050015|cpd15325	CHEBI:57970
Tetracosahexaenoic acid in MNX differ in stereochemistry layer and nisinate is synonym in HMDB	nisinate	CHEBI:77202|CHEBI:77366|HMDB02007|LMFA01030804|LMFA01030822	HMDB13025|53481586
L-octanoyl carnitine in MNX is synonym to octanoyl carnitine in HMDB. Same formula but differ in both main and stereochemical inchikey layers	octanoyl carnitine	C02838|CHEBI:18102|CHEBI:13147|CHEBI:21366|CHEBI:44613|CHEBI:6279|CHEBI:73039|CHEBI:86051|HMDB00791|HMDB00834|LMFA07070002|LMFA07070095|cpd01833	21889559
Leu-leu in MNX differ from Leucylleucine in stereochemistry layer	Leucylleucine	CHEBI:73531	HMDB28933|C11332|cpd08189|CHEBI:6418|94244|leuleu

Case in which reaction context would have provided more information than metabolite name Our search only retrieved perfect name matches for BiGG database and they don't have connectivity			
	ceramide 1-phosphate(2-)	C02960|CHEBI:16197|CHEBI:13955|CHEBI:23067|CHEBI:3548|CHEBI:57674|crmp_hs|cpd01899	CHEBI:84404|crmp_cho
	5beta-cholestane-3alpha,7alpha,27-triol	C05444|CHEBI:28540|CHEBI:1702|CHEBI:20226|HMDB12455|xol7ah3|LMST04030011|LMST04030020|LMST04030146|cpd03227	HMDB60138|CE0233
	15(R)-Hydroxy-(5Z,8Z,11Z,13E)-eicosatetraenoate	CHEBI:63989|CHEBI:78837|LMFA03060030|LMFA03060067	CE2565
	o-methylhippurate	CHEBI:68455|HMDB11723	CE2934
	3alpha,7alpha,12alpha-trihydroxy-5beta-cholestan-27-al	C01301|CHEBI:16466|CHEBI:11896|CHEBI:1696|CHEBI:20218|CHEBI:48940|CHEBI:48941|HMDB03533|HMDB06263|thcholst|LMST04030088|LMST04030108|LMST04030161|LMST04030164|cpd00955	CE4872
	5,6-indolequinone-2-carboxylate	C17938|CHEBI:81394|cpd17924	CE1562
	3-hydroxy butyryl carnitine	CHEBI:72995|CHEBI:84842|HMDB13127|LMFA07070037|LMFA07070071	3bcrn
	butyryl carnitine	C02862|CHEBI:21949|CHEBI:7676|HMDB02013|LMFA07070003|LMFA07070054|cpd01840	c4crn
	succinyl carnitine	CHEBI:73034|HMDB61717|LMFA07070101	c4dc
	isovaleryl carnitine	C20826|CHEBI:70819|CHEBI:73025|HMDB00688|LMFA07070076|LMFA07070077	ivcrn
	erythro-5-hydroxy-L-lysinium(1+)	C01211|CHEBI:51807|CHEBI:14890|CHEBI:8441|pcollg5hlys	CHEBI:58357|HMDB62570|6994839|hlys
	gums	HMDB40679	gum

In this list BiGG identifier from MNX connects only to MNX which also does not connect further to other databases. Name nomenclature there suggests they stereochemistry or protonation differences			
antipyrene in pubchem is a different compound than antipyrine	antipyrine	antipyrene	CHEBI:31225|HMDB15503|C13244|2206|cpd19134
only match with GEMap was BiGG identifier for CHO with no external information	de-Fuc form of PA6	s2l2n2m2masn	l2n2m2masn
	pristanoyl coa	pristcoa	CHEBI:77250|CHEBI:28542|HMDB02057|C07297|441253|CE5126
	beta-D-fructofuranose 2,6-bisphosphate(4-)	f26bp	CHEBI:58579|21117974|CHEBI:10374|CHEBI:22767|CHEBI:28013|CHEBI:32966|CHEBI:32967|CHEBI:32968|CHEBI:40591|CHEBI:40595|CHEBI:41014|CHEBI:42553|HMDB01058|HMDB03973|HMDB60444|C05378|fdp_B
	I-antigen	galacglcgalacglcgal14acglcgalgluside_hs	G00067|G00078|cpd21554|cpd21564|CHEBI:61610|73427349
	trimethylenediaminium	13dampp	CHEBI:57484|CHEBI:70977|4030255
	3-hydroxypropanoyl-CoA(4-)	3hpcoa	CHEBI:58528|HMDB62572|44229175
	chenodeoxycholoyl coenzyme a	dcholcoa	CHEBI:28701|HMDB06292|C05337|11966205|11953854

	dolichyl phosphate(2-)	dolp_L	CHEBI:57683
	dolichyl phosphate(2-)	dolp_U	CHEBI:57683
	protein-linked asparagine residue (N-glycosylation site)	Asn_X_Ser_Thr	cpd30173
	3alpha,7alpha,12alpha-Trihydroxy-5beta-cholestanoyl-CoA(S)	cholcoas	CHEBI:15493|HMDB02178|C05448|15942872
	Adrenyl carnitine	adrncrn	HMDB06321|53477825|LMFA07070061
	Hexadecenoyl carnitine	hdcecrn	CHEBI:17490|HMDB00222|C02990|11953816|LMFA07070004
	2-decaprenyl-5-hydroxy-6-methoxy-3-methyl-1,4-benzoquinone	2dpmhobq	CHEBI:50771|HMDB62565|25010754|LMPR02010031
	2-decaprenyl-6-methoxy-3-methyl-1,4-benzoquinone	2dp6mobq_me	CHEBI:50772|HMDB60251|LMPR02010032|25010748
	Tn antigen	Tn_antigen	C04387|G00023|cpd21520|CHEBI:53608|447272
	chondroitin sulfate E (GalNAc4,6diS-GlcA) proteoglycan	cspg_e	HMDB62462|170301
	cis-dodec-3-enoyl-CoA(4-)	dd3coa	CHEBI:58543|HMDB62648|45266682
	docosa-4,7,10,13,16-pentaenoic acid	dcsptn1	CHEBI:65136|HMDB01976|6441454|LMFA01030182|5282848
	4-oxo-retinoic acid	oretn	CHEBI:269971|HMDB06285|C16678|6437063|CHEBI:80656|LMPR01090026
	sialyl-Tn antigen	sTn_antigen	G00035|cpd21528
	2,3-bisphosphonato-D-glycerate(5-)	23dpg	CHEBI:58248
	aflatoxin B1 exo-8,9-epoxide	eaflatoxin	CHEBI:30725|C19586|HMDB06558|104756|cpd20840
	Electron transfer flavoprotein oxidized	etfox	cpd30310
	Electron transfer flavoprotein reduced	etfrd	C04570|cpd30311
	13-cis-retinoyl glucuronide	13_cis_retnglc	CHEBI:28870|HMDB03141|C11061|5281877
stereochemistry difference (HMDB S-glycerate is synonym of L-glyceric acid)	(S)-Glycerate	glyc_S	HMDB06372|6326776|CHEBI:74324|CHEBI:10999|CHEBI:12985|CHEBI:16659|CHEBI:21027|CHEBI:21030|CHEBI:22299|CHEBI:24348|CHEBI:24349|CHEBI:32398|CHEBI:33508|CHEBI:33846|CHEBI:33871|CHEBI:4187|CHEBI:41990|CHEBI:71671|HMDB00139|HMDB31818|C00258|cpd00223|glyc_R
	GQ1b	gq1b_hs	LMSP0601AV00|G00117|cpd21585
	keratan sulfate I	ksi	CHEBI:111173|HMDB62483|5087
	linoleate	lnlc	CHEBI:30245|C01595|CHEBI:17351|HMDB00673|5280450|LMFA01030120|cpd01122
	alpha-linolenate	lnlnca	CHEBI:32387|CHEBI:27432|HMDB01388|C06427|5280934|LMFA01030152|LMFA01030153|cpd03850

	PA6	s2l2fn2m2masn	G00251|cpd21626|11963565
	(S)-3-sulfonatolactate(2-)	sl__L|cpd15906	CHEBI:61289|C11499|45479474
	omega hydroxy dodecanoate (n-C12:0)	whddca	cpd30623
	omega hydroxy tetradecanoate (n-C14:0)	whttdca	cpd30627
	F1alpha	f1a	101175278
	chondroitin sulfate E (GalNAc4,6diS-GlcA), precursor 3	cs_e_pre3	HMDB62463|170301
	GlcNAc-GlcA-(Gal)2-Xyl	hs_deg24	HMDB62477|108223
	1-methylimidazole-4-acetaldehyde	3mldz	CHEBI:28104|C05827|HMDB04181|193545|cpd03458
	myo-inositol 1,3,4,5,6-pentakisphosphate(10-)	mi13456p	CHEBI:57733|CHEBI:57257|23615305
	4a-hydroxytetrahydrobiopterin	thbpt4acam	CHEBI:15374|HMDB02281|C15522|129803|CHEBI:15642|46173804|cpd03897|4hthb
	acryloyl-CoA(4-)	prpncoa	CHEBI:57367|45266595|CHEBI:13722|CHEBI:15513|CHEBI:26301|CHEBI:8488|HMDB02307|HMDB06507|C00894|LMFA07050282|LMFA07050283|cpd00663|pp2coa
	chondroitin sulfate E (GalNAc4,6diS-GlcA), precursor 5a	cs_e_pre5a	HMDB62464|10837
	keratan sulfate II (core 2-linked), degradation product 1	ksii_core2_deg1	CHEBI:50198|HMDB62484|7213
	Tachysterol 3	ts3	HMDB06560|5283713|LMST03020223
Strict name match only to BiGG database to the CHO compound homologues			
	galactosyl glucosyl ceramide	galgluside_hs	galgluside_cho
	trihexosyl ceramide	thcrm_hs	thcrm_cho
	alpha GalNAc globoside	acgagbside_hs	acgagbside_cho
	beta GalNAc globoside	acgbgbside_hs	acgbgbside_cho
	disialyl galactosylgloboside	acnacngalgbside_hs	acnacngalgbside_cho
	1-alkyl 2-acylglycerol 3-phosphocholine	ak2gchol_hs	ak2gchol_cho
	1-alkyl 2-acylglycerol 3-phosphoethanolamine	ak2gpe_hs	ak2gpe_cho
	1-alkenyl 2-acylglycerol 3-phosphoethanolamine plasmalogen	dak2gpe_hs	dak2gpe_cho
	1-alkyl 2-acylglycerol 3-phosphate	ak2gp_hs	ak2gp_cho
	1-alkyldihydroxyacetone phosphate	akdhap_hs	akdhap_cho
	O-alkylglycerone phosphate(2-)	akgp_hs	akgp_cho
	1-alkyl 2-lysoglycerol 3-phosphocholine	ak2lgchol_hs	ak2lgchol_cho
	R group 1 Coenzyme A	R1coa_hs	R1coa_cho
	R group 2 Coenzyme A	R2coa_hs	R2coa_cho

	R group 3 Coenzyme A	R3coa_hs	R3coa_cho
	R group 5 Coenzyme A	R5coa_hs	R5coa_cho
	R group 6 Coenzyme A	R6coa_hs	R6coa_cho
	GM2alpha	gm2a_hs	gm2a_cho
	(dimannosyl),(phosphoethanolaminyl)-mannosyl-glucosylaminyl-acylphosphatidylinositol (H6)	m2emgacpail_hs	m2emgacpail_cho
	deacylated-glycophosphatidylinositol (GPI)-anchored protein	dgpi_prot_hs	dgpi_prot_cho
	1 acyl phosphoglycerol	1glyc_hs	1glyc_cho
	fucosyl galactosylgloboside	fucgalgbside_hs	fucgalgbside_cho
	GD1beta	gd1b2_hs	gd1b2_cho
	9-O-Acetylated GT3	oagt3_hs	oagt3_cho
	1-alkyl 2-acteylglycerol 3-phosphocholine	paf_hs	paf_cho
	mannosyl-glucosaminyl-acylphosphatidylinositol (H2)	mgacpail_hs	mgacpail_cho
	phosphoethanolaminyl-mannosyl-glucosylaminyl-acylphosphatidylinositol (H5)	emgacpail_hs	emgacpail_cho
	dimannosyl-glucosaminyl-acylphosphatidylinositol (H3)	m2gacpail_hs	m2gacpail_cho
	trimannosyl-glucosaminyl-acylphosphatidylinositol (H4)	m3gacpail_hs	m3gacpail_cho
	phosphoethanolaminyl-trimannosyl-glucosaminyl-acylphosphatidylinositol (H6)	em3gacpail_hs	em3gacpail_cho
	(trimannosyl),(phosphoethanolaminyl)-mannosyl-glucosaminyl-acylphosphatidylinositol (M4C)	m3emgacpail_hs	m3emgacpail_cho
	sialyl galactosylgloboside	acngalgbside_hs	acngalgbside_cho
	sphingomyelin betaine	sphmyln_hs	sphmyln_cho
	GP1c alpha	gp1calpha_hs	gp1calpha_cho
	glucosaminyl-acylphosphatidylinositol	gacpail_hs	gacpail_cho

Name match to BiGG database to the CHO compound homologues, but more information obtained from name matches in other databases			
	Type IA glycolipid	acgalfucgalacglcgalgluside_hs	G00042|cpd21533|acgalfucgalacglcgalgluside_cho
	Type IIIAb	acgalfucgalacgalfucgalacglcgal14acglcgalgluside_hs	G00075|cpd21561|acgalfucgalacgalfucgalacglcgal14acglcgalgluside_cho
	3,8-LD1	acnacngal14acglcgalgluside_hs	G00064|cpd21552|acnacngal14acglcgalgluside_cho
	lysophosphatidic acid	alpa_hs	CHEBI:132742|CHEBI:16975|C00681|LMGP10050000|25163997|C03849|cpd00517|1ag3p|CHEBI:62837|HMDB07855|LMGP10050008|5311263|alpa_cho
	phosphatidate(2-)	pa_hs	CHEBI:57739|pa_cho

	D-galactosyl-N-acylsphingosine	galside_hs	CHEBI:18390|CHEBI:36498|C02686|G11121|cpd01743|galside_cho
	N-acylsphingosine	crm_hs	CHEBI:17761|CHEBI:52639|C00195|LMSP02010000|HMDB04947|cpd00167|cera-D|crm_cho
	globoside	gbside_hs	CHEBI:61360|CHEBI:18259|C03272|G00094|cpd02088|gbside_cho
	GA2	ga2_hs	CHEBI:27731|C06135|G00123|cpd03656|CHEBI:465284|C07019|3454|ga2_cho
	gibberellin A1(1-)	ga1_hs	CHEBI:58524|11877115|ga1_cho
	ganglioside GM1	gm1_hs	CHEBI:61048|9963963|gm1_cho
	GD2	gd2_hs	CHEBI:28648|C06134|G00114|cpd03655|6450346|gd2_cho
	GD1b	gd1b_hs	CHEBI:28175|C06141|LMSP0601AQ00|G00115|G02755|G04395|G04443|cpd03661|gd1b_cho
	GT2	gt2_hs	LMSP0601AO00|G00119|cpd21587|gt2_cho
	GT1c	gt1c_hs	LMSP0601AR00|G00120|cpd21588|gt1c_cho
	galactosylgloboside	galgbside_hs	G00097|cpd21579|galgbside_cho
	i-antigen	galacglcgal14acglcgalgluside_hs	G00067|G00078|cpd21554|cpd21564|CHEBI:61610|73427349|galacglcgalacglcgal14acglcgalgluside_cho
	nLc7Cer	acglc13galacglcgal14acglcgalgluside_hs	G00068|cpd21555|acglc13galacglcgal14acglcgalgluside_cho
	GM1alpha	gm1a_hs	LMSP0601BM00|CHEBI:61577|73427350|gm1a_cho
	D-glucosyl-N-acylsphingosine	gluside_hs	CHEBI:18368|C01190|HMDB00596|C03108|22833534|G10238|cpd00878|gluside_cho
	CDP-diacylglycerol(2-)	cdpdag_hs	CHEBI:58332|cdpdag_cho
	diglyceride	dag_hs	CHEBI:18035|C00165|C00641|cpd11423|12dgr_BS|dag_cho
	phosphatidylethanolamine	pe_hs	CHEBI:16038|HMDB60501|C00350|CHEBI:47767|17754131|LMGP02010000|cpd11456|pe|pe_PA|psetha_BS|pe_cho
	phosphatidylglycerol(1-)	pglyc_hs	CHEBI:60523|pglyc_cho
	triglyceride	tag_hs	CHEBI:17855|C00422|LMGL03010000|cpd11677|triglyc_SC|CHEBI:75844|HMDB05474|LMGL03010371|5322095|tag_cho
	dihydroceramide	dhcrm_hs	CHEBI:31488|C12126|HMDB06752|16755624|LMSP02020000|cpd08908|CHEBI:67033|HMDB11761|LMSP02020008|5283573|dhcrm_cho
	III3Fuc-nLc6Cer	fuc13galacglcgal14acglcgalgluside_hs	G00076|cpd21562|fuc13galacglcgal14acglcgalgluside_cho
	V3Fuc,III3Fuc-nLc6Cer	fucfuc132galacglcgal14acglcgalgluside_hs	G00090|cpd21575|fucfuc132galacglcgal14acglcgalgluside_cho
	Leb glycolipid	fucfucgalacglcgalgluside_hs	G00045|cpd21536|fucfucgalacglcgalgluside_cho
	Lacto-N-fucopentaosyl III ceramide	fucgal14acglcgalgluside_hs	G00060|cpd21548|fucgal14acglcgalgluside_cho
	GD1c	gd1c_hs	LMSP0601BN00|G00126|cpd21592|gd1c_cho
	GP1c	gp1c_hs	LMSP0601AX00|G00122|cpd21590|gp1c_cho
	GQ1balpha	gq1balpha_hs	G00129|cpd21594|gq1balpha_cho

	GT1a	gt1a_hs	CHEBI:27691|C06138|LMSP0601AW00|G00112|cpd03658|gt1a_cho
	lysophosphatidylcholine	lpchol_hs	CHEBI:60479|HMDB11128|24779491|cpd16818|lysphchol|LMGP01050043|5311264|lpchol_cho
	monoacylglycerol 2	mag_hs	CHEBI:17389|C02112|mag_cho
	9-O-Acetylated GD3	oagd3_hs	CHEBI:61730|73427362|oagd3_cho
	phosphatidylserine	ps_hs	CHEBI:18303|HMDB14291|6323481|CHEBI:11750|C02737|LMGP03010000|cpd11455|ps|ps_BS|ps_PA|ps_cho
	sphingosylphosphorylcholine	spc_hs	C03640|cpd02284|CHEBI:17689|LMSP0106000 1|9847290|spc_cho
	cholesterol ester	xolest2_hs	CHEBI:17002|C02530|cpd01664|xolest2_cho
	V3Fuc-nLc6Cer	fuc132galacglcgal14acglcgal|gluside_hs	G00089|cpd21574|fuc132galacglcgal14acglcgalguside_cho
	GD3	gd3_hs	CHEBI:28424|C06133|G00113|G05200|cpd03654|gd3_cho
	GT3	gt3_hs	LMSP0601AL00|G00118|cpd21586|CHEBI:28541|73427348|gt3_cho
	phosphatidylinositol 4,5-bisphosphate	pail45p_hs	CHEBI:18348|C04637|cpd12582|ptd1145bp_SC|5311358|pail45p_cho
	1-phosphatidyl-1D-myo-inositol 4-phosphate(3-)	pail4p_hs	CHEBI:58178|pail4p_cho
	1-phosphatidyl-1D-myo-inositol 3,4,5-trisphosphate(7-)	pail345p_hs	CHEBI:57836|pail345p_cho
	1-phosphatidyl-1D-myo-inositol 3,4-bisphosphate(5-)	pail34p_hs	CHEBI:57658|pail34p_cho
	1-phosphatidyl-1D-myo-inositol 5-phosphate(3-)	pail5p_hs	CHEBI:57795|pail5p_cho
	GM1b	gm1b_hs	G00125|cpd21591|gm1b_cho
	GT1b	gt1b_hs	CHEBI:28058|C06140|LMSP0601AT00|G00116|G01804|G04091|cpd03660|gt1b_cho
	GT1aalpha	gt1alpha_hs	G00128|cpd21593|gt1alpha_cho
	GQ1c	gq1c_hs	LMSP0601AU00|G00121|cpd21589|gq1c_cho

Table C.3.4 Statistics on reaction directionality at different stages of flux constraining for Recon 2 v4 prior to our curation of the directionalities based on thermodynamics. The model had 44% pre-assigned bidirectional reactions column 1. Flux variability analysis (FVA) of the feasible flux space imposing only the mass-balance constraints of the network reduces this proportion to 20.7% (column 2). At the same time about 21% of the reactions carry no flux. Additional constraining by imposing reaction thermodynamics on this network without condition specific metabolomics data didn't decrease significantly the flexibility of network, as shown by the thermodynamics-based flux variability analysis performed (TFVA) (column 3). Column 4 is just a statistic on the reactions from the model that have thermodynamics constraints.

Reaction directionality	*Directionalities imposed as flux constraints*	*Directionalities after evaluating the feasible solution space using only network mass-balances*	*Directionalities after evaluating the feasible solution space after applying thermodynamics constraints*	*Directionalities of reactions with thermodynamics constraints for default metabolite concentration ranges*
BI	3274 (44%)	1541 (20.7%)	1459 (19.6%)	3337 (91.1%)
F	3618 (48.6%)	3505 (47.1%)	3524 (47.4%)	313 (8.5%)
R	54 (0.7%)	280 (3.8%)	330 (4.4%)	14 (0.4%)
BO	496 (6.7%)	2116 (28.4%)	2129 (28.6%)	-

Table C.3.5 Statistics on thermodynamics curation per metabolite and reaction for two GEM networks: iMM1415 for mouse (*Mus Musculus*) and Recon 2 v4 for Homo Sapiens. Both GEMs were mapped with MetaNetX (http://www.metanetx.org) for comparison and GEMap was used to retrieve the compound structural information from the respective database links.

GEM stats	***iMM1415***		***Recon 2 v4***
Metabolites	2775		5063
Unique metabolites	1503		2626
Unique reactions	3728		7442
Reactions without transports	2559		4922
Transport reactions	1169		2520
Mapping method	Just using provided KEGG id	Using GEMap pipelines in DRAMA	Using GEMap pipelines in DRAMA
Number of metabolites with ΔfG°	1694	1811	3072
Number of unique metabolites with ΔfG°	808	865	1401
Number of reactions with ΔrG°	1094	1979	3671
Number of reactions excluding transports with ΔrG°	426	1210	2180

C.4 Supplementary Tables for Chapter 4

Table C.4.1 Detailed listing of dataset sources for normal and cancer phenotypes. In the info field, it is reported the type of experimental setup: tissue metabolomics or cell line culture for instance, as well as the condition for which the data was collected.

Normal Human	Cell Type	Phenotype	Condition	Organism	Tissue Type	Info	REF
NH1	293sf-3f6 human embryonic kidney cells	normal	normoxia	human	kidney	Uptake and secretion rates measured under exponential growth; overflow expected.	henry et al. (2011) [260]
NH2	hek293 (human embryonic kidney)	normal	normoxia	human	kidney	Metabolomics concentrations were measured at 4 different time points during exponential growth phase for parental cell. We use the data for the parental cells and give the range over the 4 time points for the metabolomics. We only use the measurements in hplc that are in fmol/cell.	dietmair et al (2012) [261]
NH3	hepatocytes	normal	normoxia	human	liver	We use the metabolomics data measured for both normal and cancer condition at the control condition (culture with ethanol)	qin et al (2013) [262]
NH4	lung fibroblasts	normal	normoxia	human	lung	Cell culture. Data was taken from the control condition; no mention of overflow in text but possible due to high lactate/glucose ratio	munger et al. (2008) [263]
NH5	hh25 (spontaneous immortalized cell line derived from normal human hepatocytes)	normal	normoxia	human	liver	Proliferative characteristics of hh25 assessed	smalley et al. (1999) [264]
NH6	surrounding tissue from colon tumor	normal	normoxia	human	colon	Metabolomics performed in tissue sample with normal control taken from surrounding tissue	hirayama et al. (2009) [265]
NH7	surrounding tissue from lung tumor	normal	normoxia	human	lung		kami et al. (2012) [266]
NH8	surrounding tissue from prostate tumor	normal	normoxia	human	prostate		kami et al. (2012) [266]
NH9	surrounding tissue from	normal	normoxia	human	stomach		hirayama et al. (2009)

	stomach tumor						(265)
NH10	surrounding tissue from prostate tumor	normal	normoxia	human	prostate		giskeodegard et al. (2013)
NH11	surrounding tissue lung human fibroblasts	normal	normoxia	human	lung	Cell growth measured for control fibroblasts taken from 3 patients undergoing lobectomy/pneumonectomy for removal of a primary lung tumor, from areas with no histologic evidence of disease	ramirez et al. (2011) (267)
NH12	surrounding tissue lung human fibroblasts	normal	normoxia	human	lung	Cell growth measured for control fibroblasts taken from 5 patients undergoing lobectomy/pneumonectomy for removal of a primary lung tumor, from areas with no histologic evidence of disease	ramos et al. (2001) (268)
Cancer Human							
CH1	a549 cell line (lung carcinoma)	cancer	hypoxia	human	lung	Cell uptake and secretion rates of glucose, lactate, glutamine and glutamate were measured under normoxia and hypoxia. Taken from graphics	metallo et al. (2012) (269)
CH2	a549 cell line (lung carcinoma)	cancer	normoxia	human	lung		metallo et al. (2012) (269)
CH3	colon tumor	cancer	normoxia	human	colon	Metabolomics performed in tissue sample with normal control taken from surrounding tissue	hirayama et al. (2009) (265)
CH4	glioblastoma cell	cancer	normoxia	human	cns	Cell culture exhibit aerobic glycolysis	deberardinis et al. (2007) (191)
CH5	hepg2 cells (ecacc wiltshire)	cancer	normoxia	human	liver	Cell viability in culture accessed	erro et al. (2013) (270)
CH6	hepg2 and h7402 cells	cancer	normoxia	human	liver		su et al. (2011) (271)
CH7	hepg2	cancer	normoxia	human	liver	Cell culture for 13C labeling experiments (fluxes and concentrations)	hofmann et al. (2007) (272)
CH8	jhh-7 cells (hepatocellular carcinoma)	cancer	normoxia	human	liver	We use the metabolomics data measured for both normal and cancer condition at the control condition (culture with ethanol)	qin et al (2013) (262)
CH9	lung tumor	cancer	normoxia	human	lung	Metabolomics performed in tissue sample with normal control taken from surrounding tissue	kami et al. (2012) (266)
CH10	786-o (renal cell adenocarcinoma)	cancer	normoxia	human	kidney	nci60 cancer cell lines	jain et al. (2012) (217)
CH11	a498 (kidney adenocarcinoma)	cancer	normoxia	human	kidney		
CH12	a549/atcc (lung	cancer	normoxia	human	lung		

	adenocarcinoma)						
CH13	achn (renal adenocarcinoma; derived from metastatic site in pleural effusion)	cancer	normoxia	human	kidney		
CH14	bt-549 (ductal carcinoma)	cancer	normoxia	human	breast		
CH15	caki-1 (clear cell carcinoma)	cancer	normoxia	human	kidney		
CH16	ccrf-cem (acute lymphoblastic leukemia (t lymphoblast))	cancer	normoxia	human	leukemia		
CH17	colo 205 (dukes type d, colorectal adenocarcinoma)	cancer	normoxia	human	colon		
CH18	du-145 (prostate carcinoma; derived from metastatic site in brain)	cancer	normoxia	human	prostate		
CH19	ekvx (lung adenocarcinoma)	cancer	normoxia	human	lung		
CH20	hcc-2998 (colon carcinoma)	cancer	normoxia	human	colon		
CH21	hct-116 (colorectal carcinoma)	cancer	normoxia	human	colon		
CH22	hct-15 (dukes type c, colorectal adenocarcinoma)	cancer	normoxia	human	colon		
CH23	hl-60(tb) (acute promyelocytic leukemia)	cancer	normoxia	human	leukemia		
CH24	hop-62 (lung adenocarcinoma)	cancer	normoxia	human	lung		
CH25	hop-92 (lung adenocarcinoma, large cell, undifferentiated)	cancer	normoxia	human	lung		
CH26	hs 578t (mammary gland carcinoma)	cancer	normoxia	human	breast		
CH27	ht29 (colorectal adenocarcinoma)	cancer	normoxia	human	colon		
CH28	grov1 (ovary cystadenocarcinoma)	cancer	normoxia	human	ovary		
CH29	k562 (chronic	cancer	normoxia	human	leukemia		

	myelogenous leukemia)						
CH30	km12 (colon adenocarcinoma, grade iii)	cancer	normoxia	human	colon		
CH31	lox imvi (malignant amelanotic melanoma)	cancer	normoxia	human	skin		
CH32	m14 (malignant melanoma)	cancer	normoxia	human	skin		
CH33	malme-3m (fibroblast malignant melanoma derived from metastatic site)	cancer	normoxia	human	lung		
CH34	mcf7 (mammary gland adenocarcinoma)	cancer	normoxia	human	breast		
CH35	mda-mb-231/atcc (mammary gland adenocarcinoma; derived from metastatic site in pleural effusion)	cancer	normoxia	human	breast		
CH36	mda-mb-435 (melanoma adenocarcinoma)	cancer	normoxia	human	skin		
CH37	mda-mb-468 (mammary gland adenocarcinoma; derived from metastatic site in pleural effusion)	cancer	normoxia	human	breast		
CH38	molt-4 (acute lymphoblast leukemia)	cancer	normoxia	human	leukemia		
CH39	nci-adr-res (ovarian adenocarcinoma)	cancer	normoxia	human	ovary		
CH40	nci-h23 (lung adenocarcinoma; non-small cell lung cancer)	cancer	normoxia	human	lung		
CH41	nci-h226 (squamous cell carcinoma; mesothelioma; derived from metastatic site in pleural effusion)	cancer	normoxia	human	lung		
CH42	nci-h322m (small cell bronchi alveolar carcinoma	cancer	normoxia	human	lung		
CH43	nci-h460 (carcinoma; large cell lung cancer)	cancer	normoxia	human	lung		

CH44	nci-h522 (lung adenocarcinoma; non-small cell lung cancer)	cancer	normoxia	human	lung		
CH45	ovcar-3 (ovary adenocarcinoma)	cancer	normoxia	human	ovary		
CH46	ovcar-4 (ovary adenocarcinoma)	cancer	normoxia	human	ovary		
CH47	ovcar-5 (ovary adenocarcinoma)	cancer	normoxia	human	ovary		
CH48	ovcar-8 (ovary adenocarcinoma)	cancer	normoxia	human	ovary		
CH49	pc-3 (prostate grade iv, adenocarcinoma; derived from metastatic site in bone)	cancer	normoxia	human	prostate		
CH50	rpmi 8226 (plasmacytoma, myeloma (b lymphocyte))	cancer	normoxia	human	leukemia		
CH51	rxf 393 (poorly differentiated hypernephroma)	cancer	normoxia	human	kidney		
CH52	sf-268 (glioblastoma, anaplastic astrocytoma)	cancer	normoxia	human	cns		
CH53	sf-295 (glioblastoma-multiform)	cancer	normoxia	human	cns		
CH54	sf-539 (glioblastoma)	cancer	normoxia	human	cns		
CH55	sk-mel-2 (malignant melanoma)	cancer	normoxia	human	skin		
CH56	sk-mel-5 (malignant melanoma)	cancer	normoxia	human	skin		
CH57	sk-mel-28 (malignant melanoma)	cancer	normoxia	human	skin		
CH58	sk-ov-3 (ovary adenocarcinoma)	cancer	normoxia	human	ovary		
CH59	sn12c (renal carcinoma)	cancer	normoxia	human	kidney		
CH60	snb-19 (glioblastoma)	cancer	normoxia	human	cns		
CH61	snb-75 (astrocytoma)	cancer	normoxia	human	cns		
CH62	sr (large cell, immunoblastic lymphoma)	cancer	normoxia	human	leukemia		
CH63	sw620 (colorectal	cancer	normoxia	human	colon		

	adenocarcinoma; derived from metastatic site in lymph node)							
CH64	t-47d (ductal carcinoma)	cancer	normoxia	human	breast	The uptake and secretion rates and biomass was measured for these cells under high and low myc gene expression (they can be tuned for the regulation of such gene). We use the high myc data as it is the oncogenic standard		murphy et al. (2013) (273)
CH65	tk-10 (spindle cell carcinoma)	cancer	normoxia	human	kidney			
CH66	u251 (glioblastoma)	cancer	normoxia	human	cns			
CH67	uacc-62 (malignant melanoma)	cancer	normoxia	human	skin			
CH68	uacc-257 (malignant melanoma)	cancer	normoxia	human	skin			
CH69	uo-31 (renal carcinoma)	cancer	normoxia	human	kidney			
CH70	p493-6 b-cell line (lymphoma)	cancer	normoxia	human	lymphatic tissues/blood cells			
CH71	prostate tumor	cancer	normoxia	human	prostate	Metabolomics performed in tissue sample with normal control taken from surrounding tissue		kami et al. (2012) (266)
CH72	stomach tumor	cancer	normoxia	human	stomach			hirayama et al. (2009) (265)
CH73	prostate tumor	cancer	normoxia	human	prostate			giskeodegard et al. (2013) (274)

Table C.4.2 Information on growth rate measurements and data extraction methodology (data point extraction from growth curves, for instance) per dataset.

Normal Human	Cell Type	Phenotype	Condition	Organism	Tissue Type	Growth Rate [h-1]	Info
NH1	293sf-3f6 human embryonic kidney cells	normal	normoxia	human	kidney	0.029	Maximum cell specific growth rate, which is similar to values previously reported for HEK-293 cells grown in bioreactor with a similar serum-free medium formulation
NH2	hek293 (human embryonic kidney)	normal	normoxia	human	kidney	0.028±0.001	Exponential growth of cell line
NH3	hepatocytes	normal	normoxia	human	liver		
NH4	lung fibroblasts	normal	normoxia	human	lung		
NH5	hh25 (spontaneous immortalized cell line derived from normal human hepatocytes)	normal	normoxia	human	liver	0.0033±0.00029	Specific growth rate computed from proliferation graphics with different substrates and averaged
NH6	surrounding tissue from colon tumor	normal	normoxia	human	colon		
NH7	surrounding tissue from lung tumor	normal	normoxia	human	lung		
NH8	surrounding tissue from prostate tumor	normal	normoxia	human	prostate		
NH9	surrounding tissue from stomach tumor	normal	normoxia	human	stomach		
NH10	surrounding tissue from prostate tumor	normal	normoxia	human	prostate		
NH11	surrounding tissue lung human fibroblasts	normal	normoxia	human	lung	0.0071±0.0004	Deviation is the gap between two cell growths and center point from Thy-1(+) and (-) cells. Data taken from graphic and used in computation of growth rate.
NH12	surrounding tissue lung human fibroblasts	normal	normoxia	human	lung	0.0069±0.0006	Deviation is the gap between two cell growths and center point from medium supplements with 1% or 10% FBS
Cancer Human							
CH1	a549 cell line (lung carcinoma)	cancer	hypoxia	human	lung	0.026	Apparent growth rate obtained from supplement graphic for A549 hypoxic cell culture
CH2	a549 cell line (lung carcinoma)	cancer	normoxia	human	lung	0.0314±0.0012	Apparent growth rate obtained from supplement graphic for A549 normoxic cell culture
CH3	colon tumor	cancer	normoxia	human	colon		

CH4	glioblastoma cell	cancer	normoxia	human	cns		
CH5	hepg2 cells (ecacc wiltshire)	cancer	normoxia	human	liver	0.0097	Specific growth rate computed from proliferation graphic
CH6	hepg2 and h7402 cells	cancer	normoxia	human	liver	0.022	Specific growth rate computed from proliferation graphic
CH7	hepg2	cancer	normoxia	human	liver		
CH8	jhh-7 cells (hepatocellular carcinoma)	cancer	normoxia	human	liver		
CH9	lung tumor	cancer	normoxia	human	lung		
CH10	786-o (renal cell adenocarcinoma)	cancer	normoxia	human	kidney	0.0446	From http://www.nexcelom.com/Applications/Cancer-Cells.html#feature1
CH11	a498 (kidney adenocarcinoma)	cancer	normoxia	human	kidney	0.0149	
CH12	a549/atcc (lung adenocarcinoma)	cancer	normoxia	human	lung	0.0436	
CH13	achn (renal adenocarcinoma; derived from metastatic site in pleural effusion)	cancer	normoxia	human	kidney	0.0363	
CH14	bt-549 (ductal carcinoma)	cancer	normoxia	human	breast	0.0185	
CH15	caki-1 (clear cell carcinoma)	cancer	normoxia	human	kidney	0.0256	
CH16	ccrf-cem (acute lymphoblastic leukemia (t lymphoblast))	cancer	normoxia	human	leukemia	0.0374	
CH17	colo 205 (dukes type d, colorectal adenocarcinoma)	cancer	normoxia	human	colon	0.042	
CH18	du-145 (prostate carcinoma; derived from metastatic site in brain)	cancer	normoxia	human	prostate	0.0309	
CH19	ekvx (lung adenocarcinoma)	cancer	normoxia	human	lung	0.0229	
CH20	hcc-2998 (colon carcinoma)	cancer	normoxia	human	colon	0.0317	
CH21	hct-116 (colorectal carcinoma)	cancer	normoxia	human	colon	0.0574	
CH22	hct-15 (dukes type c, colorectal adenocarcinoma)	cancer	normoxia	human	colon	0.0485	

CH23	hl-60(tb) (acute promyelocytic leukemia)	cancer	normoxia	human	leukemia	0.0349	
CH24	hop-62 (lung adenocarcinoma)	cancer	normoxia	human	lung	0.0256	
CH25	hop-92 (lung adenocarcinoma, large cell, undifferentiated)	cancer	normoxia	human	lung	0.0125	
CH26	hs 578t (mammary gland carcinoma)	cancer	normoxia	human	breast	0.0185	
CH27	ht29 (colorectal adenocarcinoma)	cancer	normoxia	human	colon	0.0512	
CH28	grov1 (ovary cystadenocarcinoma)	cancer	normoxia	human	ovary	0.0322	
CH29	k562 (chronic myelogenous leukemia)	cancer	normoxia	human	leukemia	0.05102	
CH30	km12 (colon adenocarcinoma, grade iii)	cancer	normoxia	human	colon	0.0422	
CH31	lox imvi (malignant amelanotic melanoma)	cancer	normoxia	human	skin	0.0487	
CH32	m14 (malignant melanoma)	cancer	normoxia	human	skin	0.038	
CH33	malme-3m (fibroblast malignant melanoma derived from metastatic site)	cancer	normoxia	human	lung	0.0216	
CH34	mcf7 (mammary gland adenocarcinoma)	cancer	normoxia	human	breast	0.0394	
CH35	mda-mb-231/atcc (mammary gland adenocarcinoma; derived from metastatic site in pleural effusion)	cancer	normoxia	human	breast	0.0238	
CH36	mda-mb-435 (melanoma adenocarcinoma)	cancer	normoxia	human	skin	0.0387	
CH37	mda-mb-468 (mammary gland adenocarcinoma; derived from metastatic site in pleural effusion)	cancer	normoxia	human	breast	0.0161	
CH38	molt-4 (acute lymphoblast leukemia)	cancer	normoxia	human	leukemia	0.0358	

CH39	nci-adr-res (ovarian adenocarcinoma)	cancer	normoxia	human	ovary	0.0294
CH40	nci-h23 (lung adenocarcinoma; non-small cell lung cancer)	cancer	normoxia	human	lung	0.0299
CH41	nci-h226 (squamous cell carcinoma; mesothelioma; derived from metastatic site in pleural effusion)	cancer	normoxia	human	lung	0.0164
CH42	nci-h322m (small cell bronchi alveolar carcinoma	cancer	normoxia	human	lung	0.0283
CH43	nci-h460 (carcinoma; large cell lung cancer)	cancer	normoxia	human	lung	0.0562
CH44	nci-h522 (lung adenocarcinoma; non-small cell lung cancer)	cancer	normoxia	human	lung	0.0262
CH45	ovcar-3 (ovary adenocarcinoma)	cancer	normoxia	human	ovary	0.0288
CH46	ovcar-4 (ovary adenocarcinoma)	cancer	normoxia	human	ovary	0.0241
CH47	ovcar-5 (ovary adenocarcinoma)	cancer	normoxia	human	ovary	0.0205
CH48	ovcar-8 (ovary adenocarcinoma)	cancer	normoxia	human	ovary	0.0383
CH49	pc-3 (prostate grade iv, adenocarcinoma; derived from metastatic site in bone)	cancer	normoxia	human	prostate	0.0369
CH50	rpmi 8226 (plasmacytoma, myeloma (b lymphocyte))	cancer	normoxia	human	leukemia	0.0298
CH51	rxf 393 (poorly differentiated hypernephroma)	cancer	normoxia	human	kidney	0.0158
CH52	sf-268 (glioblastoma, anaplastic astrocytoma)	cancer	normoxia	human	cns	0.0302
CH53	sf-295 (glioblastoma-multiform)	cancer	normoxia	human	cns	0.0338
CH54	sf-539 (glioblastoma)	cancer	normoxia	human	cns	0.0282

CH55	sk-mel-2 (malignant melanoma)	cancer	normoxia	human	skin	0.0219	
CH56	sk-mel-5 (malignant melanoma)	cancer	normoxia	human	skin	0.0396	
CH57	sk-mel-28 (malignant melanoma)	cancer	normoxia	human	skin	0.0285	
CH58	sk-ov-3 (ovary adenocarcinoma)	cancer	normoxia	human	ovary	0.0205	
CH59	sn12c (renal carcinoma)	cancer	normoxia	human	kidney	0.0338	
CH60	snb-19 (glioblastoma)	cancer	normoxia	human	cns	0.0289	
CH61	snb-75 (astrocytoma)	cancer	normoxia	human	cns	0.0159	
CH62	sr (large cell, immunoblastic lymphoma)	cancer	normoxia	human	leukemia	0.034	
CH63	sw620 (colorectal adenocarcinoma; derived from metastatic site in lymph node)	cancer	normoxia	human	colon	0.049	
CH64	t-47d (ductal carcinoma)	cancer	normoxia	human	breast	0.0219	
CH65	tk-10 (spindle cell carcinoma)	cancer	normoxia	human	kidney	0.0194	
CH66	u251 (glioblastoma)	cancer	normoxia	human	cns	0.042	
CH67	uacc-62 (malignant melanoma)	cancer	normoxia	human	skin	0.0319	
CH68	uacc-257 (malignant melanoma)	cancer	normoxia	human	skin	0.0259	
CH69	uo-31 (renal carcinoma)	cancer	normoxia	human	kidney	0.0239	
CH70	p493-6 b-cell line (lymphoma)	cancer	normoxia	human	lymphatic tissues/ blood cells	0.0293±0.0008	Specific growth rate for high Myc
CH71	prostate tumor	cancer	normoxia	human	prostate		
CH72	stomach tumor	cancer	normoxia	human	stomach		
CH73	prostate tumor	cancer	normoxia	human	prostate		

Table C.4.3 Detailed listing of information on conversion factors and assumptions made to convert flux and intracellular concentration measurements into standard units.

Normal Human	Cell Type	Tissue Type	Extracellular Fluxes	Intracellular concentrations	Info/Conversions
NH1	293sf-3f6 human embryonic kidney cells	kidney	X		Cell dry weight value 514 pg/cell from Hek293 cells in [Dietmair et al. (2012) A multi-omics analysis of recombinant protein production in Hek293 cells]
NH2	hek293 (human embryonic kidney)	kidney		X	For the volume of the cell we assume the average value for HeLa 2e-12 L [http://kirschner.med.harvard.edu/files/bionumbers/fundamentalBioNumbersHandout.pdf].
NH3	hepatocytes	liver		X	HepG2 volume of 2.85e-12 L per cell [http://bionumbers.hms.harvard.edu/bionumber.aspx?&id=104614&ver=7]
NH4	lung fibroblasts	lung	X	X	Cell dry weight assumed equal to HeLa cell 400 pg. Average cell volume used [1.92e-12, 2.88e-12] L include primary human lung fibroblast values [http://ajplung.physiology.org/content/275/5/L998]
NH5	hh25 (spontaneous immortalized cell line derived from normal human hepatocytes)	liver			Just for growth rate
NH6	surrounding tissue from colon tumor	colon		X	For conversion to volume in concentrations the assumed density of soft tissues to be 1g/mL.
NH7	surrounding tissue from lung tumor	lung		X	
NH8	surrounding tissue from prostate tumor	prostate		X	
NH9	surrounding tissue from stomach tumor	stomach		X	
NH10	surrounding tissue from prostate tumor	prostate		X	
NH11	surrounding tissue lung human fibroblasts	lung		X	
NH12	surrounding tissue lung human fibroblasts	lung		X	
Cancer Human					
CH1	a549 cell line (lung carcinoma)	lung	X		Cell dry weight assumed equal to HeLa cell 400 pg

CH2	a549 cell line (lung carcinoma)	lung	X		
CH3	colon tumor	colon		X	For conversion to volume in concentrations the assumed density of soft tissues to be 1g/mL.
CH4	glioblastoma cell	cns	X		Cell dry weight assumed equal to HeLa cell 400 pg
CH5	hepg2 cells (ecacc wiltshire)	liver			Just for growth rate
CH6	hepg2 and h7402 cells	liver			
CH7	hepg2	liver	X	X	Used human hepatocyte dry weight (431 pg) [Skorina A.D., 1983: Dry weight of isolated human hepatocytes in the normal conditions and during chronic hepatitis] Because HepG2 reference value was smaller (419 pg) [Niklas J. et al. (2009) Effects of drugs in subtoxic concentrations on the metabolic fluxes in human hepatoma cell line Hep G2]. We used a range for hepatocyte cell volume to make the bounds [1.92e-12, 2.88e-12] L (include HeLa and HepG2 cell volumes [http://kirschner.med.harvard.edu/files/bionumbers/fundamentalBioNumbersHandout.pdf][http://bionumbers.hms.harvard.edu/bionumber.aspx?&id=104614&ver=7])
CH8	jhh-7 cells (hepatocellular carcinoma)	liver		X	HepG2 volume of 2.85e-12 L per cell [http://bionumbers.hms.harvard.edu/bionumber.aspx?&id=104614&ver=7]
CH9	lung tumor	lung		X	For conversion to volume in concentrations the assumed density of soft tissues to be 1g/mL.
CH10	786-o (renal cell adenocarcinoma)	kidney	X	X	Cell dry weight assumed equal to HeLa cell 400 pg
CH11	a498 (kidney adenocarcinoma)	kidney	X	X	
CH12	a549/atcc (lung adenocarcinoma)	lung	X	X	
CH13	achn (renal adenocarcinoma; derived from metastatic site in pleural effusion)	kidney	X	X	
CH14	bt-549 (ductal carcinoma)	breast	X	X	
CH15	caki-1 (clear cell carcinoma)	kidney	X	X	
CH16	ccrf-cem (acute lymphoblastic leukemia (t lymphoblast))	leukemia	X	X	
CH17	colo 205 (dukes type d, colorectal adenocarcinoma)	colon	X	X	
CH18	du-145 (prostate carcinoma; derived from metastatic site in brain)	prostate	X	X	

CH19	ekvx (lung adenocarcinoma)	lung	X	X	
CH20	hcc-2998 (colon carcinoma)	colon	X	X	
CH21	hct-116 (colorectal carcinoma)	colon	X	X	
CH22	hct-15 (dukes type c, colorectal adenocarcinoma)	colon	X	X	
CH23	hl-60(tb) (acute promyelocytic leukemia)	leukemia	X	X	
CH24	hop-62 (lung adenocarcinoma)	lung	X	X	
CH25	hop-92 (lung adenocarcinoma, large cell, undifferentiated)	lung	X	X	
CH26	hs 578t (mammary gland carcinoma)	breast	X	X	
CH27	ht29 (colorectal adenocarcinoma)	colon	X	X	
CH28	grov1 (ovary cystadenocarcinoma)	ovary	X	X	
CH29	k562 (chronic myelogenous leukemia)	leukemia	X	X	
CH30	km12 (colon adenocarcinoma, grade iii)	colon	X	X	
CH31	lox imvi (malignant amelanotic melanoma)	skin	X	X	
CH32	m14 (malignant melanoma)	skin	X	X	
CH33	malme-3m (fibroblast malignant melanoma derived from metastatic site)	lung	X	X	
CH34	mcf7 (mammary gland adenocarcinoma)	breast	X	X	
CH35	mda-mb-231/atcc (mammary gland adenocarcinoma; derived from metastatic site in pleural effusion)	breast	X	X	
CH36	mda-mb-435 (melanoma adenocarcinoma)	skin	X	X	
CH37	mda-mb-468 (mammary gland adenocarcinoma; derived from metastatic site in pleural effusion)	breast	X	X	
CH38	molt-4 (acute lymphoblast	leukemia	X	X	

	leukemia)				
CH39	nci-adr-res (ovarian adenocarcinoma)	ovary	X	X	
CH40	nci-h23 (lung adenocarcinoma; non-small cell lung cancer)	lung	X	X	
CH41	nci-h226 (squamous cell carcinoma; mesothelioma; derived from metastatic site in pleural effusion)	lung	X	X	
CH42	nci-h322m (small cell bronchioalveolar carcinoma	lung	X	X	
CH43	nci-h460 (carcinoma; large cell lung cancer)	lung	X	X	
CH44	nci-h522 (lung adenocarcinoma; non-small cell lung cancer)	lung	X	X	
CH45	ovcar-3 (ovary adenocarcinoma)	ovary	X	X	
CH46	ovcar-4 (ovary adenocarcinoma)	ovary	X	X	
CH47	ovcar-5 (ovary adenocarcinoma)	ovary	X	X	
CH48	ovcar-8 (ovary adenocarcinoma)	ovary	X	X	
CH49	pc-3 (prostate grade iv, adenocarcinoma; derived from metastatic site in bone)	prostate	X	X	
CH50	rpmi 8226 (plasmacytoma, myeloma (b lymphocyte))	leukemia	X	X	
CH51	rxf 393 (poorly differentiated hypernephroma)	kidney	X	X	
CH52	sf-268 (glioblastoma, anaplastic astrocytoma)	cns	X	X	
CH53	sf-295 (glioblastoma-multiform)	cns	X	X	
CH54	sf-539 (glioblastoma)	cns	X	X	
CH55	sk-mel-2 (malignant melanoma)	skin	X	X	
CH56	sk-mel-5 (malignant melanoma)	skin	X	X	
CH57	sk-mel-28 (malignant	skin	X	X	

	melanoma)				
CH58	sk-ov-3 (ovary adenocarcinoma)	ovary	X	X	
CH59	sn12c (renal carcinoma)	kidney	X	X	
CH60	snb-19 (glioblastoma)	cns	X	X	
CH61	snb-75 (astrocytoma)	cns	X	X	
CH62	sr (large cell, immunoblastic lymphoma)	leukemia	X	X	
CH63	sw620 (colorectal adenocarcinoma; derived from metastatic site in lymph node)	colon	X	X	
CH64	t-47d (ductal carcinoma)	breast	X	X	
CH65	tk-10 (spindle cell carcinoma)	kidney	X	X	
CH66	u251 (glioblastoma)	cns	X	X	
CH67	uacc-62 (malignant melanoma)	skin	X	X	
CH68	uacc-257 (malignant melanoma)	skin	X	X	
CH69	uo-31 (renal carcinoma)	kidney	X	X	
CH70	p493-6 b-cell line (lymphoma)	lymphatic tissues/ blood cells	X		Cell dry weight is 150 pg as indicated in the paper
CH71	prostate tumor	prostate		X	For conversion to volume in concentrations assume density of soft tissues to be 1g/mL.
CH72	stomach tumor	stomach			Just for growth rate
CH73	prostate tumor	prostate			

Table C.4.4 List of uptakes blocked in Recon 2 v4 based on the presence of Coenzyme A (CoA), acyl carrier protein (ACP) and phosphate (P) groups in the metabolites structure. Not allowing the consumption of these molecules enforces their metabolism inside the cell.

Compound name	Reaction in model	Compound name	Reaction in model	Compound name	Reaction in model
Nicotinamide adenine dinucleotide phosphate	EX_nadp_e	1-alkyl 2-lysoglycerol 3-phosphocholine	EX_ak2lgchol_hs_e	sphingosine 1-phosphate(1-)	EX_sphs1p_e
ATP	EX_atp_e	Aquacob(III)alamin	EX_aqcobal_e	Thiamin monophosphate	EX_thmmp_e
ADP	EX_adp_e	arachidyl coenzyme A	EX_arachcoa_e	thiamine(1+) triphosphate(4-)	EX_thmtp_e
Nicotinamide adenine dinucleotide	EX_nad_e	CTP	EX_ctp_e	UMP	EX_ump_e
Coenzyme A	EX_coa_e	phosphatidylethanolamine	EX_pe_hs_e	UDPglucose	EX_udpg_e
3'',5''-cyclic GMP(1-)	EX_35cgmp_e	phosphatidylglycerol(1-)	EX_pglyc_hs_e	7,8-dihydroneopterin 3''-triphosphate(4-)	EX_ahdt_e
UDP-D-galactose(2-)	EX_udpgal_e	UTP	EX_utp_e	Pantetheine 4''-phosphate	EX_pan4p_e
UDP	EX_udp_e	CDP	EX_cdp_e	Pyridoxal 5''-phosphate	EX_pydx5p_e
Malonyl-CoA	EX_malcoa_e	dGTP	EX_dgtp_e	4-nitrophenyl phosphate(2-)	EX_HC01104_e
ITP(3-)	EX_itp_e	dGMP	EX_dgmp_e	3-oxodocosanoyl-CoA	EX_CE2250_e
IDP(3-)	EX_idp_e	dTTP	EX_dttp_e	'2'',3''-cyclic UMP(1-)'	EX_23cump_e
Flavin adenine dinucleotide oxidized	EX_fad_e	dTDP	EX_dtdp_e	'3''-UMP(2-)'	EX_3ump_e
FMN	EX_fmn_e	Dephospho-CoA	EX_dpcoa_e	1-alkyl 2-acteylglycerol 3-phosphocholine	EX_paf_hs_e
GTP	EX_gtp_e	dTMP	EX_dtmp_e	phosphatidylserine	EX_ps_hs_e
GDP	EX_gdp_e	1 acyl phosphoglycerol	EX_1glyc_hs_e	sphingosylphosphorylcholine	EX_spc_hs_e
3'',5''-cyclic AMP(1-)	EX_camp_e	ADP-D-ribose 2''-phosphate(4-)	EX_adprbp_e	(2S,3R)-2-azaniumyl-3-hydroxyoctadecyl phosphate	EX_sph1p_e
D-Glucose 1-phosphate	EX_g1p_e	lysophosphatidylcholine	EX_lpchol_hs_e	5-Phospho-alpha-D-ribose 1-diphosphate	EX_prpp_e
ADPribose	EX_adprib_e	5-Amino-1-(5-Phospho-D-ribosyl)imidazole-4-carboxamide	EX_aicar_e	CDP-ethanolamine(1-)	EX_cdpea_e
Blocking exceptions to phosphate containing compounds due to diffusion assumption: Hydrogenphosphate (EX_pi_e); Diphosphate (EX_ppi_e); Blocking exceptions to Phosphate containing compounds due to existing uptake data points: GMP (EX_gmp_e); AMP (EX_amp_e); IMP (EX_imp_e); CMP (EX_cmp_e); Dihydroxyacetone phosphate (EX_dhap_e); Adenosylcobalamin (EX_adocbl_e)					

Table C.4.5 List of essential amino acids and other metabolites found in mammalian cell minimal media culture and serum.

L-histidine	L-argininium(1+)	Taurine	Oxidized glutathione	4-pyridoxate	3-hydroxy-L-kynurenine
L-threonine	L-tyrosine	trans-4-hydroxy-L-proline	S-Adenosyl-L-homocysteine	5-Hydroxyindoleacetate	kynurenate
L-valine	L-cystine	creatine	L-homoserine	anthranilate	L-2-aminoadipate(1-)
L-methionine	L-cysteine	Adenine	CMP	bilirubin(2-)	Lactose
L-tryptophan	Ornithine	Hypoxanthine	IMP	N-carbamoyl-beta-alaninate	L-kynurenine
L-phenylalanine	(S)-lactate	Spermidine	GMP	L-Carnitine	Nicotinate
L-isoleucine	D-glucose	L-citrulline	Isocitrate	dCMP	Nicotinamide
L-leucine	O2	spermine(4+)	glycine betaine	Deoxycytidine	Nicotinate D-ribonucleotide
L-lysinium(1+)	Ammonium	L-cystathionine	Choline	N,N-dimethylglycine	Orotate
L-glutamate(1-)	Succinate	carnosine	creatinine	Deoxyuridine	Orotidine 5-phosphate
L-glutamine	2-Oxoglutarate	Uridine	acetoacetate	O-phosphonatoethanaminium(1-)	oxalate(2-)
Glycine	Citrate	Adenosine	Cytidine	Folate	R total 2 position
L-alanine	3-Phospho-D-glycerate	Inosine	D-Glycerate 2-phosphate	sn-Glycero-3-phosphocholine	taurocholate
L-serine	Phosphoenolpyruvate	quinolinate(2-)	(S)-3-aminoisobutyric acid	glycocholate	(R)-Pantothenate
L-aspartate(1-)	AMP	Glycerol	3-hydroxyanthranilate	glycochenodeoxycholate	Bicarbonate
L-asparagine	Dihydroxyacetone phosphate	Hexadecanoate (n-C16:0)	4-hydroxybenzoate	D-glucuronate	
L-proline	4-Aminobutanoate	3-(4-hydroxyphenyl)pyruvate	L-homocysteine	D-Glyceraldehyde	

Nicotinamide	Thiamin	Folate	Thiamin	guanidinoacetic acid	
Pyridoxine	Adenosylcobalamin	calcium(2+)	potassium	Chloride	
Riboflavin	myo-inositol	sulfate	Sodium	hydrogenphosphate	
Thymine	3,3,5-triiodo-L-thyronine	Urate	xanthosine	acetate	
Thymidine	UDP-D-glucuronate	Xanthine	sulfate	linoleate	
L-thyroxine	uracil	Xanthosine 5-phosphate	hydrogenphosphate	alpha-linolenate	
Choline	Biotin	Taurodeoxycholic acid	serotonin(1+)	(R)-Pantothenate	
Diphosphate	Chloride	Reduced glutathione	Sucrose	propionate	

Table C.4.6 List of subsystems from Recon 2 v4 that are not explicitly kept in the *RedHuman* core network.

Alkaloid synthesis	*N-glycan degradation*
Androgen and estrogen synthesis and metabolism	*N-glycan synthesis*
Arachidonic acid metabolism	*Nucleotide salvage pathway*
Biotin metabolism	*Nucleotide sugar metabolism*
Blood group synthesis	*O-glycan synthesis*
C5-branched dibasic acid metabolism	*Phenylalanine metabolism*
Chondroitin sulfate degradation	*Phosphatidylinositol phosphate metabolism*
Chondroitin synthesis	*Selenoamino acid metabolism*
CoA catabolism	*Starch and sucrose metabolism*
CoA synthesis	*Steroid metabolism*
Cytochrome metabolism	*Stilbene, coumarine and lignin synthesis*
Dietary fiber binding	*Taurine and hypotaurine metabolism*
Galactose metabolism	*Tetrahydrobiopterin metabolism*
Glycosphingolipid metabolism	*Thiamine metabolism*
Glyoxylate and dicarboxylate metabolism	*Transport, golgi apparatus*
Heme degradation	*Ubiquinone synthesis*
Heme synthesis	*Vitamin A metabolism*
Heparan sulfate degradation	*Vitamin B12 metabolism*
Inositol phosphate metabolism	*Vitamin B2 metabolism*
Keratan sulfate degradation	*Vitamin B6 metabolism*
Keratan sulfate synthesis	*Vitamin C metabolism*
Limonene and pinene degradation	*Vitamin D metabolism*
Linoleate metabolism	*Vitamin E metabolism*
Lipoate metabolism	*Xenobiotics metabolism*
beta-Alanine metabolism	

Table C.4.7 List of metabolites that have lost extracellular reactions in *RedHuman* (green), that have been completely eliminated in all compartments of *RedHuman* (yellow), and that have been removed in some compartments of *RedHuman* (typically l, n, g, x, r), while still existing in others (e, c, m) (orange). Compounds that have extracellular reactions removed but are still present in the network in the extracellular media (e) are generated by extracellular reactions using other compounds. Notation for compartments (compart.): c (cytosol) , e (extracellular), n (nucleus), m (mitochondria), r (endoplasmic reticulum), g (Golgi), x (peroxisome), l (lysosome).

Observation	Metabolite name	Removed in Compart.	Exist in Compart.
Extracellular reaction removed despite data, but no removal of intracellular compound	Dihydroxyacetone phosphate	e	c,x,m
	Orotate	-	c,e
	serotonin(1+)	-	c,e
	taurocholate	-	c,x,e
	L-thyroxine	-	c,e
	Sucrose	e	no other compart. in GEM
	Taurodeoxycholic acid	e	no other compart. In GEM
	glycochenodeoxycholate	-	e,c,x
Extracellular reaction and intracellular compound removed despite data	creatine	e,c,m	-
	Hypoxanthine	e,l,c,x	-
	carnosine	e,c,l	-
	IMP	e,c,m	-
	GMP	e,g,c,m,n,l	-
	Cytidine	e,l,m,n,l	-
	Glycerol	e,c,m	-
	3-(4-hydroxyphenyl)pyruvate	e,c,m	-
	(S)-3-aminoisobutyric acid	e,c,m	-
	4-pyridoxate	e,c	-
	Deoxycytidine	e,c,m,n,l	-
	Lactose	e,l,g,c	-
	Nicotinate	e,c	-
	D-glucitol	e,c	-
	Thiamin	e,c,m	-
	Thymine	e,c,m	-
	Thymidine	e,c,l,m	-
	Urate	e,c,x,n	-
Extracellular reaction and intracellular compound in some compart. removed despite data	Inosine	m,l	c,e
	CMP	g,m,n,l,r,e	c
	bilirubin(2-)	r	c,e
	D-glucuronate	l,e	c,r
	oxalate(2-)	x	c,m,e
	3,3',5-triiodo-L-thyronine	r	c,e
	UMP	g,l,m,n,e	c,r
Intracellular metabolites completely removed despite data	tyraminium	c	-
	L-cystathionine	c,m	-
	Guanosine	e,m,c,l	-
	D-gluconate	c	-
	dTMP	c,l,m,e,n	-
	UDPglucose	c,r,g,e	-
	3-hydroxyanthranilate	c	-
	4-hydroxybenzoate	m	-
	5-Hydroxyindoleacetate	c,m	-

Observation	Metabolite name	Removed in Compart.	Exist in Compart.
Intracellular metabolites removed in some compartments despite data	D-glucose	g,l	e,c,r
	L-glutamine	l	c,e,m
	Glycine	l	m,x,c,e
	L-alanine	l	m,x,e,c
	L-aspartate(1-)	l	c,m,e
	L-serine	g,l,r	c,e,m,x
	L-asparagine	l	c,e,m
	L-proline	l,r	e,m,c
	L-histidine	l	e,c,m
	L-threonine	l,m	c,e
	L-argininium(1+)	l	m,e,c
	L-tyrosine	l,m	e,c
	L-valine	l	e,c,m
	L-methionine	l,m	c,e
	L-tryptophan	l	e,c
	L-phenylalanine	l,m	e,c
	L-isoleucine	l	e,c,m
	L-leucine	l	e,c,m
	L-lysinium(1+)	l,n,x	c,m,e
	Succinate	r,x	c,e,m
	(S)-malate(2-)	x	c,m,e
	2-Oxoglutarate	r	m,c,e,x
	ATP	g,l,n	c,m,r,e,x
	L-cysteine	l,m	c,e
	3-Phospho-D-glycerate	m	c
	alpha-D-Ribose 5-phosphate	m	c,r
	ADP	e,g,l,n	c,m,x,r
	AMP	g,l	c,m,x,r,e
	GTP	e,n	c,m
	GDP	e,g,n	c,m
	UTP	m,n	c,e
	Putrescine	x	m,c,e
	beta-alanine	m	c,e
	4-Aminobutanoate	l	c,m,e,n
	Taurine	x	e,c
	trans-4-hydroxy-L-proline	r	c,m,e
	Adenine	l	e,c
	Spermidine	x	c,e
	spermine(4+)	x	c,e
	Uridine	l,m,n	c,e
	Adenosine	l,m	c,e
	Reduced glutathione	r,x	c,e,m
	S-Adenosyl-L-homocysteine	n	c,r,m,e
	S-Adenosyl-L-methionine	n	c,r,m
	Nicotinamide	l,n	m,c,r,x

					adenine dinucleotide phosphate - reduced		
	N-carbamoyl-beta-alaninate	c	-		Coenzyme A	g,l,n	m,c,x,r,e
	choline phosphate(1-)	c,g,l	-		CDP	e,m,n	c
	O-phosphonatoethanaminium(1-)	c,r	-		Acetyl-CoA	g,n,r	m,c,x
	3-hydroxy-L-kynurenine	c,m	-		Malonyl-CoA	e,l,m,n,r,x	c
	kynurenate	c,m	-		CTP	e,m,n	c
	L-kynurenine	c,m	-		Nicotinamide adenine dinucleotide	e,n	x,c,m,r
	Xanthine	c,x	-		Nicotinamide adenine dinucleotide phosphate	e,l,n	m,c,r,x
	Xanthosine 5'-phosphate	c	-		Flavin adenine dinucleotide oxidized	e,x	m,r,c
	xanthosine	c	-		UDP-N-acetyl-alpha-D-glucosamine(2-)	g,r	c
	etha	c	-		UDP	e,g,l,m,n,r	c
	sn-Glycero-3-phosphocholine	c	-		(R)-3-hydroxybutyrate	m	e,c
	D-Glyceraldehyde	c,m	-		Choline	g,n,r	e,c,m
	guanidinoacetic acid	c	-		2-deoxyadenosine	l	c,e
	Nicotinate D-ribonucleotide	c,n,m	-		dCMP	l,m,n	c
	Orotidine 5'-phosphate	c	-		Deoxyuridine	m,n	c,e
	D-Glucose 1-phosphate	c,e	-		Folate	m	e,c
	(S)-2-Aminobutanoate	c	-		(R)-Pantothenate	m	e,c
	creatinine	c	-		propionate	m,x	c,e
	quinolinate(2-)	c	-		UDP-D-glucuronate	g	c,r
					myo-inositol	g,r	c,e
					UMP	e,g,l,m,n	c,r
					L-thyroxine	r	e,c

Table C.4.8 The 15 lumped reactions added to the core network of *RedHuman* to generate a network that is capable of producing all biomass building blocks. The components of the biomass reaction originally in Recon 2 v4 are present in the ***Biomass components*** columns and if a lumped reaction was generated it is indicated with ✔, followed by the respective lumped reaction.

Biomass components		***Lumped reaction***
Substrates		
Water		
ATP	✔	proton + hydrogenphosphate + 2-deoxyadenosine <=> 2-Deoxy-D-ribose 1-phosphate + Adenine
L-glutamate(1-)		
L-aspartate(1-)		
GTP	✔	ATP + 5-Phospho-alpha-D-ribose 1-diphosphate + Guanine <=> ADP + Diphosphate + GDP
L-asparagine		
L-alanine		
L-cysteine		
L-glutamine		
Glycine		
L-serine		
L-threonine		
L-lysinium(1+)		
L-argininium(1+)		
L-methionine		
1-phosphatidyl-1D-myo-inositol(1-)	✔	ATP + myo-inositol + diglyceride <=> Water + ADP + 1-phosphatidyl-1D-myo-inositol(1-)
CTP	✔	2 proton + 5-Phospho-alpha-D-ribose 1-diphosphate + N-Carbamoyl-L-aspartate + Orotate <=> Water + carbon dioxide + Diphosphate + (S)-dihydroorotate + UMP
Phosphatidylcholine	✔	Ammonium + hydrogenphosphate + acetaldehyde + 3 S-Adenosyl-L-methionine + CTP + diglyceride + phosphatidylethanolamine <=> Water + Diphosphate + CMP + 3 S-Adenosyl-L-homocysteine + 3 proton + phosphatidylethanolamine + Phosphatidylcholine
phosphatidylethanolamine	✔	proton + ATP + L-serine + diglyceride + phosphatidylserine <=> Water + ADP + carbon dioxide + phosphatidylethanolamine + phosphatidylserine
cholesterol		
phosphatidylglycerol(1-)	✔	Water + R total 2 coenzyme A + lysophosphatidic acid + CTP + Glycerol 3-phosphate <=> hydrogenphosphate + Coenzyme A + Diphosphate + CMP + phosphatidylglycerol(1-)
cardiolipin	✔	Water + 2 ATP + 2 CTP + 2 diglyceride + Glycerol 3-phosphate <=> 2 ADP + 2 proton + hydrogenphosphate + 2 Diphosphate + 2 CMP + cardiolipin
UTP		
dGTP	✔	ATP + proton + hydrogenphosphate + Deoxyguanosine <=> Water + ADP + dGDP
dCTP	✔	4.8583 2-Deoxy-D-ribose 1-phosphate + 3.8583 uracil + Uridine + 11.25 dCDP <=> 3.8583 proton + 3.8583 hydrogenphosphate + 5.625 dCMP + 5.625 dCTP + 4.8583 Deoxyuridine + alpha-D-Ribose 1-phosphate
dATP	✔	ATP + 2-deoxyadenosine <=> ADP + dAMP(2-)
dTTP	✔	Water + 2.451 ATP + 2.2105 2-Deoxy-D-ribose 1-phosphate + 2.451 Nicotinamide adenine dinucleotide + cytosine + 1.2105 uracil + 2.451 dUMP + 2.451 5,10-Methylenetetrahydrofolate <=> Ammonium + 2.451 ADP + 3.6614 proton + 2.2105 hydrogenphosphate + 2.451 Nicotinamide adenine dinucleotide - reduced + 2.2105 Deoxyuridine + 2.451 dTDP + 2.451 Folate
D-Glucose 6-phosphate		
L-histidine		
L-tyrosine	✔	O2 + proton + Nicotinamide adenine dinucleotide - reduced + L-phenylalanine <=> Water + Nicotinamide adenine dinucleotide + L-tyrosine
L-isoleucine		
L-leucine		
L-tryptophan		
L-phenylalanine		
L-proline		
phosphatidylserine	✔	ATP + L-serine + diglyceride <=> Water + ADP + phosphatidylserine
sphingomyelin betaine	✔	O2 + proton + hydrogenphosphate + Nicotinamide adenine dinucleotide - reduced + CTP + Choline + dihydroceramide <=> 3 Water + Nicotinamide adenine dinucleotide + Diphosphate + CMP + sphingomyelin betaine
L-valine		
Products		
ADP		
proton		
hydrogenphosphate		

Bibliography

1. Grzmil, M., and B.A. Hemmings. 2012. Translation regulation as a therapeutic target in cancer. Cancer Research. 72: 3891–3900.

2. Scheper, G.C., M.S. van der Knaap, and C.G. Proud. 2007. Translation matters: protein synthesis defects in inherited disease. Nat Rev Genet. 8: 711–723.

3. Rötig, A. 2011. Human diseases with impaired mitochondrial protein synthesis. Biochimica et Biophysica Acta (BBA) - Bioenergetics. 1807: 1198–1205.

4. Boczonadi, V., and R. Horvath. 2014. Mitochondria: Impaired mitochondrial translation in human disease. The International Journal of Biochemistry & Cell Biology. 48: 77–84.

5. Wurm, F.M. 2004. Production of recombinant protein therapeutics in cultivated mammalian cells. Nat Biotechnol. 22: 1393–1398.

6. Mardinoglu, A., R. Agren, C. Kampf, A. Asplund, M. Uhlén, and J. Nielsen. 2014. Genome-scale metabolic modelling of hepatocytes reveals serine deficiency in patients with non-alcoholic fatty liver disease. Nat Comms. 5: 3083.

7. Gatto, F., H. Miess, A. Schulze, and J. Nielsen. 2015. Flux balance analysis predicts essential genes in clear cell renal cell carcinoma metabolism. Sci Rep. 5: 1029.

8. Zielinski, D.C., N. Jamshidi, A.J. Corbett, A. Bordbar, A. Thomas, and B.Ø. Palsson. 2017. Systems biology analysis of drivers underlying hallmarks of cancer cell metabolism. Sci Rep. 7: 41241.

9. Folger, O., L. Jerby, C. Frezza, E. Gottlieb, E. Ruppin, and T. Shlomi. 2011. Predicting selective drug targets in cancer through metabolic networks. Mol Syst Biol. 7: 501–501.

10. Duarte, N.C., S.A. Becker, N. Jamshidi, I. Thiele, M.L. Mo, T.D. Vo, R. Srivas, and B.Ø. Palsson. 2007. Global reconstruction of the human metabolic network based on genomic and bibliomic data. Proc Natl Acad Sci USA. 104: 1777–1782.

11. Thiele, I., N. Swainston, R.M.T. Fleming, A. Hoppe, S. Sahoo, M.K. Aurich, H. Haraldsdottir, M.L. Mo, O. Rolfsson, M.D. Stobbe, S.G. Thorleifsson, R. Agren, C. Bölling, S. Bordel, A.K. Chavali, P. Dobson, W.B. Dunn, L. Endler, D. Hala, M. Hucka, D. Hull, D. Jameson, N. Jamshidi, J.J. Jonsson, N. Juty, S. Keating, I. Nookaew, N. Le Novère, N. Malys, A. Mazein, J.A. Papin, N.D. Price, E. Selkov, M.I.

Sigurdsson, E. Simeonidis, N. Sonnenschein, K. Smallbone, A. Sorokin, J.H.G.M. van Beek, D. Weichart, I. Goryanin, J. Nielsen, H.V. Westerhoff, D.B. Kell, P. Mendes, and B.Ø. Palsson. 2013. A community-driven global reconstruction of human metabolism. Nat Biotechnol. 31: 419–425.

12. Mardinoglu, A., R. Agren, C. Kampf, A. Asplund, I. Nookaew, P. Jacobson, A.J. Walley, P. Froguel, L.M. Carlsson, M. Uhlén, and J. Nielsen. 2013. Integration of clinical data with a genome-scale metabolic model of the human adipocyte. Mol Syst Biol. 9: 649–649.

13. Becker, S.A., and B.Ø. Palsson. 2008. Context-specific metabolic networks are consistent with experiments. PLoS Comput Biol. 4: e1000082.

14. Colijn, C., A. Brandes, J. Zucker, D.S. Lun, B. Weiner, M.R. Farhat, T.-Y. Cheng, D.B. Moody, M. Murray, and J.E. Galagan. 2009. Interpreting expression data with metabolic flux models: predicting Mycobacterium tuberculosis mycolic acid production. PLoS Comput Biol. 5: e1000489.

15. Zur, H., E. Ruppin, and T. Shlomi. 2010. iMAT: an integrative metabolic analysis tool. Bioinformatics. 26: 3140–3142.

16. Wiback, S.J., R. Mahadevan, and B.Ø. Palsson. 2004. Using metabolic flux data to further constrain the metabolic solution space and predict internal flux patterns: the Escherichia coli spectrum. Biotechnol. Bioeng. 86: 317–331.

17. Cakir, T., K.R. Patil, Z.I. Onsan, K.O. Ulgen, B. Kirdar, and J. Nielsen. 2006. Integration of metabolome data with metabolic networks reveals reporter reactions. Mol Syst Biol. 2: 50.

18. Agren, R., S. Bordel, A. Mardinoglu, N. Pornputtapong, I. Nookaew, and J. Nielsen. 2012. Reconstruction of genome-scale active metabolic networks for 69 human cell types and 16 cancer types using INIT. PLoS Comput Biol. 8: e1002518.

19. Sørensen, M.A., and S. Pedersen. 1991. Absolute in vivo translation rates of individual codons in Escherichia coli. The two glutamic acid codons GAA and GAG are translated with a threefold difference in rate. Journal of Molecular Biology. 222: 265–280.

20. Varenne, S., J. Buc, R. Lloubes, and C. Lazdunski. 1984. Translation is a non-uniform process. Effect of tRNA availability on the rate of elongation of nascent polypeptide chains. Journal of Molecular Biology. 180: 549–576.

21. Curran, J.F., and M. Yarus. 1989. Rates of aminoacyl-tRNA selection at 29 sense codons in vivo. Journal of Molecular Biology. 209: 65–77.

22. Komar, A.A. 2009. A pause for thought along the co-translational folding pathway. Trends in Biochemical Sciences. 34: 16–24.

23. Hartl, U.F., and M. Hayer-Hartl. 2002. Molecular Chaperones in the Cytosol: from Nascent Chain to Folded Protein. Science. 295: 18521858–1858.

24. Braakman, I., and N.J. Bulleid. 2011. Protein folding and modification in the mammalian endoplasmic reticulum. Annu. Rev. Biochem. 80: 71–99.

25. Bonekamp, F., H. Dalbøge, and T. Christensen. 1989. Translation rates of individual codons are not correlated with tRNA abundances or with frequencies of utilization in Escherichia coli.

26. Crombie, T., J.P. Boyle, J.R. Coggins, and A.J. Brown. 1994. The Folding of the Bifunctional TRP3 Protein in Yeast is Influenced by a Translational Pause which Lies in a Region of Structural Divergence with Escherichia coli Indoleglycerol-Phosphate Synthase. Eur J Biochem. 226: 657664–664.

27. Komar, A.A. 1999. Synonymous codon substitutions affect ribosome traffic and protein folding during in vitro translation. FEBS Lett. 462: 387–391.

28. Komar, A.A., and R. Jaenicke. 1995. Kinetics of translation of gamma B crystallin and its circularly permutated variant in an in vitro cell-free system: possible relations to codon distribution and protein folding. FEBS Lett. 376: 195–198.

29. Oxender, D.L., G. Zurawski, and C. Yanofsky. 1979. Attenuation in the Escherichia coli tryptophan operon: role of RNA secondary structure involving the tryptophan codon region. Proc Natl Acad Sci USA. 76: 5524–5528.

30. Jacks, T., H.D. Madhani, F.R. Masiarz, and H.E. Varmus. 1988. Signals for ribosomal frameshifting in the Rous sarcoma virus gag-pol region. Cell. 55: 447–458.

31. P.J. Farabaugh. 1990. Ribosomal frameshifting in the yeast retrotransposon Ty: tRNAs induce slippage on a 7 nucleotide minimal site. Cell. 62: 339–352.

32. Sundararajan, A., W.A. Michaud, Q. Qian, G. Stahl, and P.J. Farabaugh. 1999. Near-cognate peptidyl-tRNAs promote +1 programmed translational frameshifting in yeast. Mol. Cell. 4: 1005–1015.

33. Proshkin, S., R.A. Rahmouni, A. Mironov, and E. Nudler. 2010. Cooperation between translating ribosomes and RNA polymerase in transcription elongation. Science. 328: 504–508.

34. Kapp, L.D., and J.R. Lorsch. 2004. The molecular mechanics of Eukaryotic translation. Annu. Rev. Biochem. 73: 657704–704.

35. Wilson, D.N., and J.H. Cate. 2012. The structure and function of the eukaryotic ribosome. Cold Spring Harbor Perspectives in Biology. 4: a011536–a011536.

36. Myasnikov, A.G., A. Simonetti, S. Marzi, and B.P. Klaholz. 2009. Structure-function insights into prokaryotic and eukaryotic translation initiation. Current Opinion in Structural Biology. 19: 300–309.

37. Rodnina, M.V., and W. Wintermeyer. 2009. Recent mechanistic insights into eukaryotic ribosomes. Current Opinion in Cell Biology. 21: 435–443.

38. Alksne, L.E., R.A. Anthony, S.W. Liebman, and J.R. Warner. 1993. An accuracy center in the ribosome conserved over 2 billion years. Proc Natl Acad Sci USA. 90: 95389541–9541.

39. MacDonald, C.T., J.H. Gibbs, and A.C. Pipkin. 1968. Kinetics of biopolymerization on nucleic acid templates. Biopolymers. 6: 1–25.

40. MacDonald,} {CT, and J.H. Gibbs. 1969. Concerning the kinetics of polypeptide synthesis on polyribosomes. Biopolymers. 7: 707–725.

41. Heinrich, R., and T.A. Rapoport. 1980. Mathematical modelling of translation of mRNA in eucaryotes; steady state, time-dependent processes and application to reticulocytes. Journal of Theoretical Biology. 86: 279–313.

42. Mehra, A., and V. Hatzimanikatis. 2006. An Algorithmic Framework for Genome-Wide Modeling and Analysis of Translation Networks. Biophysical Journal. 90: 11361146–1146.

43. Gilchrist, M.A., and A. Wagner. 2006. A model of protein translation including codon bias, nonsense errors, and ribosome recycling. Journal of Theoretical Biology. 239: 417434–434.

44. Basu, A., and D. Chowdhury. 2007. Traffic of interacting ribosomes: Effects of single-machine mechanochemistry on protein synthesis. Phys Rev E. 75: 021902.

45. Siwiak, M., and P. Zielenkiewicz. 2010. A comprehensive, quantitative, and genome-wide model of translation. PLoS Comput Biol. 6: e1000865.

46. Reuveni, S., I. Meilijson, M. Kupiec, E. Ruppin, and T. Tuller. 2011. Genome-scale analysis of translation elongation with a ribosome flow model. PLoS Comput Biol. 7: e1002127.

47. Rodnina, M.V., T. Pape, R. Fricke, L. Kuhn, and W. Wintermeyer. 1996. Initial Binding of the Elongation Factor Tu·GTP·Aminoacyl-tRNA Complex Preceding Codon Recognition on the Ribosome. Journal of Biological Chemistry. 271: 646–652.

48. Pape, T., W. Wintermeyer, and M.V. Rodnina. 1998. Complete kinetic mechanism of elongation factor Tu-dependent binding of aminoacyl-tRNA to the A site of the E.coli ribosome. EMBO J. 17: 74907497–7497.

49. Gromadski, K.B., and M.V. Rodnina. 2004. Kinetic Determinants of High-Fidelity tRNA Discrimination on the Ribosome. Mol. Cell. 13: 191–200.

50. Zouridis, H., and V. Hatzimanikatis. 2008. A Model for Protein Translation: Polysome Self-Organization Leads to Maximum Protein Synthesis Rates. Biophysical Journal. 92: 717–730.

51. Zouridis, H., and V. Hatzimanikatis. 2008. Effects of codon distributions and

tRNA competition on protein translation. Biophysical Journal. 95: 1018–1033.

52. Savelsbergh, A., V.I. Katunin, D. Mohr, F. Peske, M.V. Rodnina, and W. Wintermeyer. 2004. An Elongation Factor G-Induced Ribosome Rearrangement Precedes tRNA-mRNA Translocation. Mol. Cell. 11: 1517–1523.

53. Wohlgemuth, I., C. Pohl, J. Mittelstaet, A.L. Konevega, and M.V. Rodnina. 2011. Evolutionary optimization of speed and accuracy of decoding on the ribosome. Philosophical Transactions of the Royal Society B: Biological Sciences. 366: 2979–2986.

54. Rodnina, M.V., K.B. Gromadski, U. Kothe, and H.-J. Wieden. 2004. Recognition and selection of tRNA in translation. FEBS Lett. 579: 938–942.

55. Edelmann, P. 1977. Mistranslation in E. coli. Cell. 10: 131–137.

56. Precup, J., A.K. Ulrich, O. Roopnarine, and J. Parker. 1989. Context specific misreading of phenylalanine codons. Mol. Gen. Genet. 218: 397–401.

57. Racle, J., J. Overney, and V. Hatzimanikatis. 2012. A computational framework for the design of optimal protein synthesis. Biotechnol. Bioeng. 109: 2127–2133.

58. Dennis, P.P., and H. Bremer. 1996. Modulation of chemical composition and other parameters of the cell by growth rate. In: Neidhardt FC, editor. Escherichia coli and Salmonella. Washington, DC: ASM Press. pp. 1553–1569.

59. Dong, H., L. Nilsson, and C.G. Kurland. 1996. Co-variation of tRNA Abundance and Codon Usage inEscherichia coliat Different Growth Rates. Journal of Molecular Biology. 260: 649–663.

60. Spencer, P.S., E.'.N. Siller, J.F. Anderson, and J.M. Barral. 2012. Silent substitutions predictably alter translation elongation rates and protein folding efficiencies. Journal of Molecular Biology. 422: 328–335.

61. Rodnina, M.V., and W. Wintermeyer. 2003. Fidelity of aminoacyl-trna selection on the ribosome: Kinetic and Structural Mechanisms. Annu. Rev. Biochem. 70: 415–435.

62. Milon, P., and A.L. Konevega. 2007. Transient kinetics, fluorescence, and FRET in studies of initiation of translation in bacteria. Meth. Enzymol. 430: 1–30.

63. Petrov, A., J. Chen, S. O'Leary, A. Tsai, and J.D. Puglisi. 2012. Single-molecule analysis of translational dynamics. Cold Spring Harbor Perspectives in Biology. 4: a011551–a011551.

64. Fluitt, A., E. Pienaar, and H. Viljoen. 2007. Ribosome kinetics and aa-tRNA competition determine rate and fidelity of peptide synthesis. Computational Biology and Chemistry. 31: 335346–346.

65. Zhang, G., I. Fedyunin, O. Miekley, A. Valleriani, A. Moura, and Z. Ignatova. 2010.

Global and local depletion of ternary complex limits translational elongation. Nucleic Acids Research. 38: 4778–4787.

66. Rosenblum, G., C. Chen, J. Kaur, X. Cui, H. Zhang, H. Asahara, S. Chong, Z. Smilansky, Y.E. Goldman, and B.S. Cooperman. 2013. Quantifying elongation rhythm during full-length protein synthesis. J. Am. Chem. Soc. 135: 11322–11329.

67. Chu, D., and T. von der Haar. 2012. The architecture of eukaryotic translation. Nucleic Acids Research. 40: 10098–10106.

68. Shah, P., Y. Ding, M. Niemczyk, G. Kudla, and J.B. Plotkin. 2013. Rate-Limiting Steps in Yeast Protein Translation. Cell. 153: 1589–1601.

69. Peske, F., A. Savelsbergh, V.I. Katunin, M.V. Rodnina, and W. Wintermeyer. 2004. Conformational Changes of the Small Ribosomal Subunit During Elongation Factor G-dependent tRNA–mRNA Translocation. Journal of Molecular Biology. 343: 11831194–1194.

70. Kothe, U., and M.V. Rodnina. 2007. Codon Reading by tRNAAla with Modified Uridine in the Wobble Position. Mol. Cell. 25: 167–174.

71. Pan, D., S.V. Kirillov, and B.S. Cooperman. 2007. Kinetically competent intermediates in the translocation step of protein synthesis. Mol. Cell. 25: 519–529.

72. Cai, M. Bullard, L. Thompson, and L. Spremulli. 2000. 'Interaction of Mitochondrial Elongation Factor Tu with Aminoacyl-tRNA and Elongation Factor Ts. Journal of Biological Chemistry. 275: 20308–20314.

73. Mohanty, B.K., V.F. Maples, and S.R. Kushner. 2012. Polyadenylation helps regulate functional tRNA levels in Escherichia coli. Nucleic Acids Research. 40: 4589–4603.

74. Hentzen, D., P. Mandel, and J.P. Garel. 1972. Relation between aminoacyl-tRNA stability and the fixed amino acid. Biochim. Biophys. Acta. 281: 228–232.

75. Gillespie, D.T. 1976. A general method for numerically simulating the stochastic time evolution of coupled chemical reactions.

76. Racle, J., A.J. Stefaniuk, and V. Hatzimanikatis. 2015. Noise analysis of genome-scale protein synthesis using a discrete computational model of translation. J Chem Phys. 143: 044109.

77. Kubitschek, H.E., and J.A. Friske. 1986. Determination of bacterial cell volume with the Coulter Counter. Journal of Bacteriology. 168: 1466–1467.

78. Yu, J., J. Xiao, X. Ren, K. Lao, and X.S. Xie. 2006. Probing gene expression in live cells, one protein molecule at a time. Science. 311: 1600–1603.

79. Taniguchi, Y., P.J. Choi, G.-W.W. Li, H. Chen, M. Babu, J. Hearn, A. Emili, and X.S. Xie. 2010. Quantifying E. coli proteome and transcriptome with single-molecule sensitivity in single cells. Science. 329: 533–538.

80. Li, G.-W.W., D. Burkhardt, C. Gross, and J.S. Weissman. 2014. Quantifying absolute protein synthesis rates reveals principles underlying allocation of cellular resources. Cell. 157: 624–635.

81. Zhou, J., and K.E. Rudd. 2013. EcoGene 3.0. Nucleic Acids Research. 41: D613–24.

82. Chan, P.P., and T.M. Lowe. 2009. GtRNAdb: a database of transfer RNA genes detected in genomic sequence. Nucleic Acids Research. 37: D93–7.

83. Racle, J., F. Picard, L. Girbal, M. Cocaign-Bousquet, and V. Hatzimanikatis. 2013. A genome-scale integration and analysis of Lactococcus lactis translation data. PLoS Comput Biol. 9: e1003240.

84. Arava, Y., Y. Wang, J.D. Storey, C. Liu, P.O. Brown, and D. Herschlag. 2003. Genome-wide analysis of mRNA translation profiles in Saccharomyces cerevisiae. Proc Natl Acad Sci USA. 100: 3889–3894.

85. Oh, E., A.H. Becker, A. Sandikci, D. Huber, R. Chaba, F. Gloge, R.J. Nichols, A. Typas, C.A. Gross, G. Kramer, J.S. Weissman, and B. Bukau. 2011. Selective ribosome profiling reveals the cotranslational chaperone action of trigger factor in vivo. Cell. 147: 1295–1308.

86. Li, G.-W., E. Oh, and J.S. Weissman. 2012. The anti-Shine-Dalgarno sequence drives translational pausing and codon choice in bacteria. Nature. 484: 538–541.

87. Kazazian, H.H., and M.L. Freedman. 1968. The characterization of separated alpha and beta-chain polyribosomes in rabbit reticulocytes. Journal of Biological Chemistry. 243: 6446–6450.

88. Rose, J.K. 1977. Nucleotide sequences of ribosome recgonition sites in messenger RNAs of vesicular stomatitis virus. Proc Natl Acad Sci USA. 74: 3672–3676.

89. Revel, M., and Y. Groner. 1978. Post-Transcriptional and Translational Controls of Gene Expression in Eukaryotes. Annu. Rev. Biochem. 47: 1079–1126.

90. Sørensen, M.A., and S. Pedersen. 1998. Determination of the peptide elongation rate in vivo. Protein Synthesis. 77: 129–142.

91. Kanaya, S., Y. Yamada, Y. Kudo, and T. Ikemura. 1999. Studies of codon usage and tRNA genes of 18 unicellular organisms and quantification of Bacillus subtilis tRNAs: gene expression level and species-specific diversity of codon usage based on multivariate analysis. Gene. 238: 143155–155.

92. Cognat, V., J.-M. Deragon, E. Vinogradova, T. Salinas, C. Remacle, and L. Maréchal-

Drouard. 2008. On the evolution and expression of Chlamydomonas reinhardtii nucleus-encoded transfer RNA genes. Genetics. 179: 113–123.

93. Saltelli, A. 2002. Making best use of model evaluations to compute sensitivity indices. Computer Physics Communications. 145: 280–297.

94. Rudorf, S., M. Thommen, M.V. Rodnina, and R. Lipowsky. 2014. Deducing the kinetics of protein synthesis in vivo from the transition rates measured in vitro. PLoS Comput Biol. 10: e1003909.

95. Johansson, M., M. Lovmar, and M. Ehrenberg. 2008. Rate and accuracy of bacterial protein synthesis revisited. Curr. Opin. Microbiol. 11: 141–147.

96. Khazaie, K., J.H. Buchanan, and R.F. Rosenberger. 1984. The accuracy of Q beta RNA translation. 1. Errors during the synthesis of Q beta proteins by intact Escherichia coli cells. Eur J Biochem. 144: 485–489.

97. Bouadloun, F., D. Donner, and C.G. Kurland. 1983. Codon-specific missense errors in vivo. EMBO J. 2: 1351–1356.

98. Rice, J.B., R.T. Libby, and J.N. Reeve. 1984. Mistranslation of the mRNA encoding bacteriophage T7 0.3 protein. Journal of Biological Chemistry. 259: 6505–6510.

99. Parker, J., and G. Holtz. 1984. Control of basal-level codon misreading in Escherichia coli. Biochemical and Biophysical Research Communications. 121: 487–492.

100. Forchhammer, J., and L. Lindahl. 1971. Growth rate of polypeptide chains as a function of the cell growth rate in a mutant of Escherichia coli 15. Journal of Molecular Biology. 55: 563–568.

101. Liang, S.T., Y.C. Xu, P. Dennis, and H. Bremer. 2000. mRNA composition and control of bacterial gene expression. Journal of Bacteriology. 182: 3037–3044.

102. Gromadski, K.B., T. Schümmer, A. Strømgaard, C.R. Knudsen, T. Kinzy, and M.V. Rodnina. 2007. Kinetics of the Interactions between Yeast Elongation Factors 1A and 1Bα, Guanine Nucleotides, and Aminoacyl-tRNA. Journal of Biological Chemistry. 282: 35629–35637.

103. Segrè, D., D. Vitkup, and G.M. Church. 2002. Analysis of optimality in natural and perturbed metabolic networks. Proc Natl Acad Sci USA. 99: 15112–15117.

104. Kumar, V.S., and C.D. Maranas. 2009. GrowMatch: an automated method for reconciling in silico/in vivo growth predictions. PLoS Comput Biol. 5: e1000308.

105. Papp, B., R.A. Notebaart, and C. Pál. 2011. Systems-biology approaches for predicting genomic evolution. Nat Rev Genet. 12: 591–602.

106. Suthers, P.F., A. Zomorrodi, and C.D. Maranas. 2009. Genome-scale gene/reaction essentiality and synthetic lethality analysis. Mol Syst Biol. 5:

1295.

107. Pratapa, A., S. Balachandran, and K. Raman. 2015. Fast-SL: an efficient algorithm to identify synthetic lethal sets in metabolic networks. Bioinformatics. 31: 3299–3305.

108. Burgard, A.P., P. Pharkya, and C.D. Maranas. 2003. Optknock: a bilevel programming framework for identifying gene knockout strategies for microbial strain optimization. Biotechnol. Bioeng. 84: 647–657.

109. Chowdhury, A., A.R. Zomorrodi, and C.D. Maranas. 2014. k-OptForce: integrating kinetics with flux balance analysis for strain design. PLoS Comput Biol. 10: e1003487.

110. Park, J.M., H.M. Park, W.J. Kim, H.U. Kim, T.Y. Kim, and S.Y. Lee. 2012. Flux variability scanning based on enforced objective flux for identifying gene amplification targets. BMC Syst Biol. 6: 106.

111. Pharkya, P., and C.D. Maranas. 2006. An optimization framework for identifying reaction activation/inhibition or elimination candidates for overproduction in microbial systems. Metabolic Engineering. 8: 1–13.

112. Pharkya, P. 2004. OptStrain: A computational framework for redesign of microbial production systems. Genome Research. 14: 2367–2376.

113. Ren, S., B. Zeng, and X. Qian. 2013. Adaptive bi-level programming for optimal gene knockouts for targeted overproduction under phenotypic constraints. BMC Bioinformatics. 14: S17.

114. Andreozzi, S., A. Chakrabarti, K.C. Soh, A. Burgard, T.H. Yang, S. Van Dien, L. Miskovic, and V. Hatzimanikatis. 2016. Identification of metabolic engineering targets for the enhancement of 1,4-butanediol production in recombinant E. coli using large-scale kinetic models. Metabolic Engineering. 35: 148–159.

115. Becker, S.A., A.M. Feist, M.L. Mo, G. Hannum, B.Ø. Palsson, and M.J. Herrgard. 2007. Quantitative prediction of cellular metabolism with constraint-based models: the COBRA Toolbox. Nature Protocols. 2: 727–738.

116. Schellenberger, J., R. Que, R.M.T. Fleming, I. Thiele, J.D. Orth, A.M. Feist, D.C. Zielinski, A. Bordbar, N.E. Lewis, S. Rahmanian, J. Kang, D.R. Hyduke, and B.Ø. Palsson. 2011. Quantitative prediction of cellular metabolism with constraint-based models: the COBRA Toolbox v2.0. Nature Protocols. 6: 1290–1307.

117. Bosi, E., J.M. Monk, R.K. Aziz, M. Fondi, V. Nizet, and B.Ø. Palsson. 2016. Comparative genome-scale modelling of Staphylococcus aureus strains identifies strain-specific metabolic capabilities linked to pathogenicity. Proc. Natl. Acad. Sci. U.S.A. 113: E3801–9.

118. Jamshidi, N., and B.Ø. Palsson. 2007. Investigating the metabolic capabilities of Mycobacterium tuberculosis H37Rv using the in silico strain iNJ661 and

proposing alternative drug targets. BMC Syst Biol. 1: 26.

119. Raghunathan, A., J. Reed, S. Shin, B. Palsson, and S. Daefler. 2009. Constraint-based analysis of metabolic capacity of Salmonella typhimurium during host-pathogen interaction. BMC Syst Biol. 3: 38.

120. Bordbar, A., N.E. Lewis, J. Schellenberger, B.Ø. Palsson, and N. Jamshidi. 2010. Insight into human alveolar macrophage and M. tuberculosis interactions via metabolic reconstructions. Mol Syst Biol. 6: 422.

121. Shoaie, S., F. Karlsson, A. Mardinoglu, I. Nookaew, S. Bordel, and J. Nielsen. 2013. Understanding the interactions between bacteria in the human gut through metabolic modeling. Sci Rep. 3: 2532.

122. Heinken, A., S. Sahoo, R.M.T. Fleming, and I. Thiele. 2014. Systems-level characterization of a host-microbe metabolic symbiosis in the mammalian gut. Gut Microbes. 4: 28–40.

123. Hyötyläinen, T., L. Jerby, E.M. Petäjä, I. Mattila, S. Jäntti, P. Auvinen, A. Gastaldelli, H. Yki-Järvinen, E. Ruppin, and M. Oresic. 2016. Genome-scale study reveals reduced metabolic adaptability in patients with non-alcoholic fatty liver disease. Nat Comms. 7: 8994.

124. Väremo, L., C. Scheele, C. Broholm, A. Mardinoglu, C. Kampf, A. Asplund, I. Nookaew, M. Uhlén, B.K. Pedersen, and J. Nielsen. 2015. Proteome- and Transcriptome-Driven Reconstruction of the Human Myocyte Metabolic Network and Its Use for Identification of Markers for Diabetes. Cell Reports. 11: 921–933.

125. Resendis-Antonio, O., A. Checa, and S. Encarnación. 2010. Modeling Core Metabolism in Cancer Cells: Surveying the Topology Underlying the Warburg Effect. PLoS ONE. 5: e12383.

126. Shlomi, T., T. Benyamini, E. Gottlieb, R. Sharan, and E. Ruppin. 2011. Genome-Scale Metabolic Modeling Elucidates the Role of Proliferative Adaptation in Causing the Warburg Effect. PLoS Comput Biol. 7: e1002018.

127. Förster, J., I. Famili, P. Fu, B.Ø. Palsson, and J. Nielsen. 2003. Genome-scale reconstruction of the Saccharomyces cerevisiae metabolic network. Genome Research. 13: 244–253.

128. Babaei, P., S.-A. Marashi, and S. Asad. 2015. Genome-scale reconstruction of the metabolic network in Pseudomonas stutzeri A1501. Mol. BioSyst. 11: 3022–3032.

129. Thiele, I., D.R. Hyduke, B. Steeb, G. Fankam, D.K. Allen, S. Bazzani, P. Charusanti, F.-C. Chen, R.M.T. Fleming, C.A. Hsiung, S.C.J. De Keersmaecker, Y.-C. Liao, K. Marchal, M.L. Mo, E. Özdemir, A. Raghunathan, J.L. Reed, S.-I. Shin, S. Sigurbjörnsdóttir, J. Steinmann, S. Sudarsan, N. Swainston, I.M. Thijs, K. Zengler,

B.Ø. Palsson, J.N. Adkins, and D. Bumann. 2011. A community effort towards a knowledge-base and mathematical model of the human pathogen Salmonella Typhimurium LT2. BMC Syst Biol. 5: 8.

130. Lee, D.-S., H. Burd, J. Liu, E. Almaas, O. Wiest, A.-L. Barabási, Z.N. Oltvai, and V. Kapatral. 2009. Comparative genome-scale metabolic reconstruction and flux balance analysis of multiple Staphylococcus aureus genomes identify novel antimicrobial drug targets. Journal of Bacteriology. 191: 4015–4024.

131. Feist, A.M., C.S. Henry, J.L. Reed, M. Krummenacker, A.R. Joyce, P.D. Karp, L.J. Broadbelt, V. Hatzimanikatis, and B.Ø. Palsson. 2007. A genome-scale metabolic reconstruction for Escherichia coli K-12 MG1655 that accounts for 1260 ORFs and thermodynamic information. Mol Syst Biol. 3: 121.

132. Henry, C.S., J.F. Zinner, M.P. Cohoon, and R.L. Stevens. 2009. iBsu1103: a new genome-scale metabolic model of Bacillus subtilis based on SEED annotations. Genome Biol. 10: R69.

133. Tymoshenko, S., R.D. Oppenheim, R. Agren, J. Nielsen, D. Soldati-Favre, and V. Hatzimanikatis. 2015. Metabolic Needs and Capabilities of Toxoplasma gondii through Combined Computational and Experimental Analysis. PLoS Comput Biol. 11: e1004261.

134. Chiappino-Pepe, A., S. Tymoshenko, M. Ataman, D. Soldati-Favre, and V. Hatzimanikatis. 2017. Bioenergetics-based modeling of Plasmodium falciparum metabolism reveals its essential genes, nutritional requirements, and thermodynamic bottlenecks. PLoS Comput Biol. 13: e1005397.

135. Kanehisa, M. 1996. Toward pathway engineering: a new database of genetic and molecular pathways. Science & Technology Japan.

136. Kanehisa, M., M. Furumichi, M. Tanabe, Y. Sato, and K. Morishima. 2017. KEGG: new perspectives on genomes, pathways, diseases and drugs. Nucleic Acids Research. 45: D353–D361.

137. Maglott, D., J. Ostell, K.D. Pruitt, and T. Tatusova. 2011. Entrez Gene: gene-centered information at NCBI. Nucleic Acids Research. 39: D52–7.

138. Ma, H., A. Sorokin, A. Mazein, A. Selkov, E. Selkov, O. Demin, and I. Goryanin. 2007. The Edinburgh human metabolic network reconstruction and its functional analysis. Mol Syst Biol. 3: 135.

139. Gille, C., C. Bölling, A. Hoppe, S. Bulik, S. Hoffmann, K. Hübner, A. Karlstädt, R. Ganeshan, M. König, K. Rother, M. Weidlich, J. Behre, and H.-G. Holzhütter. 2010. HepatoNet1: a comprehensive metabolic reconstruction of the human hepatocyte for the analysis of liver physiology. Mol Syst Biol. 6: 411.

140. Fabregat, A., K. Sidiropoulos, P. Garapati, M. Gillespie, K. Hausmann, R. Haw, B. Jassal, S. Jupe, F. Korninger, S. McKay, L. Matthews, B. May, M. Milacic, K.

Rothfels, V. Shamovsky, M. Webber, J. Weiser, M. Williams, G. Wu, L. Stein, H. Hermjakob, and P. D'Eustachio. 2016. The Reactome pathway Knowledgebase. Nucleic Acids Research. 44: D481–7.

141. Thiele, I., and B.Ø. Palsson. 2010. A protocol for generating a high-quality genome-scale metabolic reconstruction. Nature Protocols. 5: 93–121.

142. Henry, C.S., M. DeJongh, A.A. Best, P.M. Frybarger, B. Linsay, and R.L. Stevens. 2010. High-throughput generation, optimization and analysis of genome-scale metabolic models. Nat Biotechnol. 28: 977–982.

143. Agren, R., L. Liu, S. Shoaie, W. Vongsangnak, I. Nookaew, and J. Nielsen. 2013. The RAVEN Toolbox and Its Use for Generating a Genome-scale Metabolic Model for Penicillium chrysogenum. PLoS Comput Biol. 9: e1002980.

144. Chaouiya, C., D. Bérenguier, S.M. Keating, A. Naldi, M.P. van Iersel, N. Rodriguez, A. Dräger, F. Büchel, T. Cokelaer, B. Kowal, B. Wicks, E. Gonçalves, J. Dorier, M. Page, P.T. Monteiro, A. von Kamp, I. Xenarios, H. de Jong, M. Hucka, S. Klamt, D. Thieffry, N. Le Novère, J. Saez-Rodriguez, and T. Helikar. 2013. SBML qualitative models: a model representation format and infrastructure to foster interactions between qualitative modelling formalisms and tools. BMC Syst Biol. 7: 135.

145. Schellenberger, J., J.O. Park, T.M. Conrad, and B.Ø. Palsson. 2010. BiGG: a Biochemical Genetic and Genomic knowledgebase of large scale metabolic reconstructions. BMC Bioinformatics. 11: 213.

146. Lang, M., M. Stelzer, and D. Schomburg. 2011. BKM-react, an integrated biochemical reaction database. BMC Biochemistry. 12: 42.

147. Placzek, S., I. Schomburg, A. Chang, L. Jeske, M. Ulbrich, J. Tillack, and D. Schomburg. 2017. BRENDA in 2017: new perspectives and new tools in BRENDA. Nucleic Acids Research. 45: D380–D388.

148. Caspi, R., R. Billington, L. Ferrer, H. Foerster, C.A. Fulcher, I.M. Keseler, A. Kothari, M. Krummenacker, M. Latendresse, L.A. Mueller, Q. Ong, S. Paley, P. Subhraveti, D.S. Weaver, and P.D. Karp. 2016. The MetaCyc database of metabolic pathways and enzymes and the BioCyc collection of pathway/genome databases. Nucleic Acids Research. 44: D471–80.

149. Wittig, U., R. Kania, M. Golebiewski, M. Rey, L. Shi, L. Jong, E. Algaa, A. Weidemann, H. Sauer-Danzwith, S. Mir, O. Krebs, M. Bittkowski, E. Wetsch, I. Rojas, and W. Müller. 2012. SABIO-RK--database for biochemical reaction kinetics. Nucleic Acids Research. 40: D790–6.

150. Kumar, A., P.F. Suthers, and C.D. Maranas. 2012. MetRxn: a knowledgebase of metabolites and reactions spanning metabolic models and databases. BMC Bioinformatics. 13: 6.

151. Hastings, J., G. Owen, A. Dekker, M. Ennis, N. Kale, V. Muthukrishnan, S. Turner,

N. Swainston, P. Mendes, and C. Steinbeck. 2016. ChEBI in 2016: Improved services and an expanding collection of metabolites. Nucleic Acids Research. 44: D1214–9.

152. Wishart, D.S., T. Jewison, A.C. Guo, M. Wilson, C. Knox, Y. Liu, Y. Djoumbou, R. Mandal, F. Aziat, E. Dong, S. Bouatra, I. Sinelnikov, D. Arndt, J. Xia, P. Liu, F. Yallou, T. Bjorndahl, R. Perez-Pineiro, R. Eisner, F. Allen, V. Neveu, R. Greiner, and A. Scalbert. 2013. HMDB 3.0--The Human Metabolome Database in 2013. Nucleic Acids Research. 41: D801–7.

153. Moretti, S., O. Martin, T. Van Du Tran, A. Bridge, A. Morgat, and M. Pagni. 2016. MetaNetX/MNXref – reconciliation of metabolites and biochemical reactions to bring together genome-scale metabolic networks. Nucleic Acids Research. 44: D523–D526.

154. Fahy, E., M. Sud, D. Cotter, and S. Subramaniam. 2007. LIPID MAPS online tools for lipid research. Nucleic Acids Research. 35: W606–12.

155. Reitz, M., O. Sacher, A. Tarkhov, D. Trumbach, and J. Gasteiger. 2004. Enabling the exploration of biochemical pathways. Org. Biomol. Chem. 2: 3226–3237.

156. Morgat, A., T. Lombardot, K.B. Axelsen, L. Aimo, A. Niknejad, N. Hyka-Nouspikel, E. Coudert, M. Pozzato, M. Pagni, S. Moretti, S. Rosanoff, J. Onwubiko, L. Bougueleret, I. Xenarios, N. Redaschi, and A. Bridge. 2017. Updates in Rhea - an expert curated resource of biochemical reactions. Nucleic Acids Research. 45: D415–D418.

157. Morgat, A., E. Coissac, E. Coudert, K.B. Axelsen, G. Keller, A. Bairoch, A. Bridge, L. Bougueleret, I. Xenarios, and A. Viari. 2012. UniPathway: a resource for the exploration and annotation of metabolic pathways. Nucleic Acids Research. 40: D761–9.

158. Mednis, M., Z. Rove, and V. Galvanauskas. 2012. ModeRator - a software tool for comparison of stoichiometric models. IEEE. pp. 97–100.

159. Qi, X., Z.M. Ozsoyoglu, and G. Ozsoyoglu. 2014. Matching metabolites and reactions in different metabolic networks. Methods. 69: 282–297.

160. Yamaguchi, A., Y. Yamamoto, J.-D. Kim, T. Takagi, and A. Yonezawa. 2012. Discriminative application of string similarity methods to chemical and non-chemical names for biomedical abbreviation clustering. BMC Genomics. 13 Suppl 3: S8.

161. ChemAxon (http://www.chemaxon.com). Molconvert (compound structure convertion). : Academic Licence.

162. Jankowski, M.D., C.S. Henry, L.J. Broadbelt, and V. Hatzimanikatis. 2008. Group Contribution Method for Thermodynamic Analysis of Complex Metabolic Networks. Biophysical Journal. 95: 1487–1499.

163. Lazar, T. 2003. Molecular Driving Forces: Statistical Thermodynamics in Chemistry and Biology. By K. A. Dill, S. Bromberg. Macromolecular Chemistry and Physics. 204: 1800–1800.

164. Henry, C.S., L.J. Broadbelt, and V. Hatzimanikatis. 2007. Thermodynamics-based metabolic flux analysis. Biophysical Journal. 92: 1792–1805.

165. Sigurdsson, M.I., N. Jamshidi, E. Steingrimsson, I. Thiele, and B.Ø. Palsson. 2010. A detailed genome-wide reconstruction of mouse metabolism based on human Recon 1. BMC Syst Biol. 4: 140.

166. Lodish, H. 2004. Molecular Cell Biology. Macmillan.

167. Alberty, R.A. 2005. Thermodynamics of Biochemical Reactions. John Wiley & Sons.

168. Casey, J.R., S. Grinstein, and J. Orlowski. 2010. Sensors and regulators of intracellular pH. Nat Rev Mol Cell Biol. 11: 50–61.

169. Haraldsdóttir, H.S., I. Thiele, and R.M.T. Fleming. 2012. Quantitative assignment of reaction directionality in a multicompartmental human metabolic reconstruction. Biophysical Journal. 102: 1703–1711.

170. Karp, G. 2009. Cell and Molecular Biology. John Wiley & Sons.

171. Cameron, R.H. 1998. Comprehensive Human Physiology: From Cellular Mechanisms to Integration.R. Greger , U. Windhorst. The Quarterly Review of Biology. 73: 112–113.

172. Bernard, T., A. Bridge, A. Morgat, S. Moretti, I. Xenarios, and M. Pagni. 2014. Reconciliation of metabolites and biochemical reactions for metabolic networks. Briefings in Bioinformatics. 15: 123–135.

173. Damerau, F.J. 1964. A technique for computer detection and correction of spelling errors. Communications of the ACM. 7: 171–176.

174. Akhondi, S.A., J.A. Kors, and S. Muresan. 2012. Consistency of systematic chemical identifiers within and between small-molecule databases. J Cheminform. 4: 35.

175. Hanahan, D., and R.A. Weinberg. 2011. Hallmarks of cancer: the next generation. Cell. 144: 646–674.

176. Hammad, N., M. Rosas-Lemus, S. Uribe-Carvajal, M. Rigoulet, and A. Devin. 2016. The Crabtree and Warburg effects: Do metabolite-induced regulations participate in their induction? Biochim. Biophys. Acta. 1857: 1139–1146.

177. Diaz-Ruiz, R., M. Rigoulet, and A. Devin. 2011. The Warburg and Crabtree effects: On the origin of cancer cell energy metabolism and of yeast glucose repression. Biochim. Biophys. Acta. 1807: 568–576.

178. Cairns, R.A., I.S. Harris, and T.W. Mak. 2011. Regulation of cancer cell metabolism. Nat. Rev. Cancer. 11: 85–95.

179. WARBURG, O. 1956. On the origin of cancer cells. Science. 123: 309–314.

180. Munyon, W.H., and D.J. Merchant. 1959. The relation between glucose utilization, lactic acid production and utilization and the growth cycle of L strain fibroblasts. Experimental Cell Research. 17: 490–498.

181. Hedeskov, C.J. 1968. Early effects of phytohaemagglutinin on glucose metabolism of normal human lymphocytes. Biochemical Journal. 110: 373–380.

182. WANG, T., C. MARQUARDT, and J. FOKER. 1976. Aerobic glycolysis during lymphocyte proliferation. Nature. 261: 702–705.

183. Brand, K., J.F. Williams, and M.J. Weidemann. 1984. Glucose and glutamine metabolism in rat thymocytes. Biochemical Journal. 221: 471–475.

184. Zhivotovsky, B., and S. Orrenius. 2009. The Warburg Effect returns to the cancer stage. Semin. Cancer Biol. 19: 1–3.

185. Moreno-Sánchez, R., S. Rodríguez-Enríquez, A. Marín-Hernández, and E. Saavedra. 2007. Energy metabolism in tumor cells. FEBS J. 274: 1393–1418.

186. Hu, J., J.W. Locasale, J.H. Bielas, J. O'Sullivan, K. Sheahan, L.C. Cantley, M.G. Vander Heiden, and D. Vitkup. 2013. Heterogeneity of tumor-induced gene expression changes in the human metabolic network. Nat Biotechnol. 31: 522–529.

187. GUPPY, M., P. LEEDMAN, X. ZU, and V. RUSSELL. 2002. Contribution by different fuels and metabolic pathways to the total ATP turnover of proliferating MCF-7 breast cancer cells. Biochemical Journal. 364: 309–315.

188. Pasdois, P., C. Deveaud, P. Voisin, V. Bouchaud, M. Rigoulet, and B. Beauvoit. 2003. Contribution of the phosphorylable complex I in the growth phase-dependent respiration of C6 glioma cells in vitro. J. Bioenerg. Biomembr. 35: 439–450.

189. Martin, M., B. Beauvoit, P.J. Voisin, P. Canioni, B. Guérin, and M. Rigoulet. 1998. Energetic and morphological plasticity of C6 glioma cells grown on 3-D support; effect of transient glutamine deprivation. J. Bioenerg. Biomembr. 30: 565–578.

190. Rodríguez-Enríquez, S., P.A. Vital-González, F.L. Flores-Rodríguez, A. Marín-Hernández, L. Ruiz-Azuara, and R. Moreno-Sánchez. 2006. Control of cellular proliferation by modulation of oxidative phosphorylation in human and rodent fast-growing tumor cells. Toxicol. Appl. Pharmacol. 215: 208–217.

191. DeBerardinis, R.J., A. Mancuso, E. Daikhin, I. Nissim, M. Yudkoff, S. Wehrli, and C.B. Thompson. 2007. Beyond aerobic glycolysis: transformed cells can engage in glutamine metabolism that exceeds the requirement for protein and

nucleotide synthesis. Proc. Natl. Acad. Sci. U.S.A. 104: 19345–19350.

192. Yang, C., J. Sudderth, T. Dang, R.M. Bachoo, R.G. Bachoo, J.G. McDonald, and R.J. DeBerardinis. 2009. Glioblastoma cells require glutamate dehydrogenase to survive impairments of glucose metabolism or Akt signaling. Cancer Research. 69: 7986–7993.

193. Le, A., A.N. Lane, M. Hamaker, S. Bose, A. Gouw, J. Barbi, T. Tsukamoto, C.J. Rojas, B.S. Slusher, H. Zhang, L.J. Zimmerman, D.C. Liebler, R.J.C. Slebos, P.K. Lorkiewicz, R.M. Higashi, T.W.M. Fan, and C.V. Dang. 2012. Glucose-independent glutamine metabolism via TCA cycling for proliferation and survival in B cells. Cell Metab. 15: 110–121.

194. Cheng, T., J. Sudderth, C. Yang, A.R. Mullen, E.S. Jin, J.M. Matés, and R.J. DeBerardinis. 2011. Pyruvate carboxylase is required for glutamine-independent growth of tumor cells. Proc. Natl. Acad. Sci. U.S.A. 108: 8674–8679.

195. Birsoy, K., R. Possemato, F.K. Lorbeer, E.C. Bayraktar, P. Thiru, B. Yucel, T. Wang, W.W. Chen, C.B. Clish, and D.M. Sabatini. 2014. Metabolic determinants of cancer cell sensitivity to glucose limitation and biguanides. Nature. 508: 108–112.

196. Caro, P., A.U. Kishan, E. Norberg, I.A. Stanley, B. Chapuy, S.B. Ficarro, K. Polak, D. Tondera, J. Gounarides, H. Yin, F. Zhou, M.R. Green, L. Chen, S. Monti, J.A. Marto, M.A. Shipp, and N.N. Danial. 2012. Metabolic signatures uncover distinct targets in molecular subsets of diffuse large B cell lymphoma. Cancer Cell. 22: 547–560.

197. Smolková, K., N. Bellance, F. Scandurra, E. Génot, E. Gnaiger, L. Plecitá-Hlavatá, P. Jezek, and R. Rossignol. 2010. Mitochondrial bioenergetic adaptations of breast cancer cells to aglycemia and hypoxia. J. Bioenerg. Biomembr. 42: 55–67.

198. Li, L., X. Zhou, W.-K. Ching, and P. Wang. 2010. Predicting enzyme targets for cancer drugs by profiling human metabolic reactions in NCI-60 cell lines. BMC Bioinformatics. 11: 501.

199. Vazquez, A., J. Liu, Y. Zhou, and Z.N. Oltvai. 2010. Catabolic efficiency of aerobic glycolysis: the Warburg effect revisited. BMC Syst Biol. 4: 58.

200. Wang, Y., J.A. Eddy, and N.D. Price. 2012. Reconstruction of genome-scale metabolic models for 126 human tissues using mCADRE. BMC Syst Biol. 6: 153.

201. Goldstein, I., K. Yizhak, S. Madar, N. Goldfinger, E. Ruppin, and V. Rotter. 2013. p53 promotes the expression of gluconeogenesis-related genes and enhances hepatic glucose production. Cancer & Metabolism. 1: 9.

202. Hensley, C.T., B. Faubert, Q. Yuan, N. Lev-Cohain, E. Jin, J. Kim, L. Jiang, B. Ko, R. Skelton, L. Loudat, M. Wodzak, C. Klimko, E. McMillan, Y. Butt, M. Ni, D. Oliver, J. Torrealba, C.R. Malloy, K. Kernstine, R.E. Lenkinski, and R.J. DeBerardinis. 2016. Metabolic Heterogeneity in Human Lung Tumors. Cell. 164: 681–694.

203. Erdrich, P., R. Steuer, and S. Klamt. 2015. An algorithm for the reduction of

genome-scale metabolic network models to meaningful core models. BMC Syst Biol. 9: 48.

204. Quek, L.-E., S. Dietmair, M. Hanscho, V.S. Martínez, N. Borth, and L.K. Nielsen. 2014. Reducing Recon 2 for steady-state flux analysis of HEK cell culture. J. Biotechnol. 184: 172–178.

205. Ataman, M., D.F. Hernandez Gardiol, G. Fengos, and V. Hatzimanikatis. 2017. redGem: Systematic Reduction and Analysis of Genome-scale Metabolic Reconstructions for Development of Consistent Core Metabolic Models. : 1–47.

206. Ataman, M., and V. Hatzimanikatis. 2017. lumGem: Systematic Generation of Subnetworks and Elementally Balanced Lumped Reactions for the Biosynthesis of Target Metabolites. : 1–42.

207. Feijó Delgado, F., N. Cermak, V.C. Hecht, S. Son, Y. Li, S.M. Knudsen, S. Olcum, J.M. Higgins, J. Chen, W.H. Grover, and S.R. Manalis. 2013. Intracellular water exchange for measuring the dry mass, water mass and changes in chemical composition of living cells. PLoS ONE. 8: e67590.

208. Kilburn, D.G., M.D. Lilly, and F.C. Webb. 1969. The energetics of mammalian cell growth. J. Cell. Sci. 4: 645–654.

209. Daugherty, M., B. Polanuyer, M. Farrell, M. Scholle, A. Lykidis, V. de Crécy-Lagard, and A. Osterman. 2002. Complete reconstitution of the human coenzyme A biosynthetic pathway via comparative genomics. Journal of Biological Chemistry. 277: 21431–21439.

210. Psychogios, N., D.D. Hau, J. Peng, A.C. Guo, R. Mandal, S. Bouatra, I. Sinelnikov, R. Krishnamurthy, R. Eisner, B. Gautam, N. Young, J. Xia, C. Knox, E. Dong, P. Huang, Z. Hollander, T.L. Pedersen, S.R. Smith, F. Bamforth, R. Greiner, B. McManus, J.W. Newman, T. Goodfriend, and D.S. Wishart. 2011. The human serum metabolome. PLoS ONE. 6: e16957.

211. Krebs, H.A. 1950. Chemical Composition of Blood Plasma and Serum. Annu. Rev. Biochem. 19: 409–430.

212. Tardito, S., A. Oudin, S.U. Ahmed, F. Fack, O. Keunen, L. Zheng, H. Miletic, P.Ø. Sakariassen, A. Weinstock, A. Wagner, S.L. Lindsay, A.K. Hock, S.C. Barnett, E. Ruppin, S.H. Mørkve, M. Lund-Johansen, A.J. Chalmers, R. Bjerkvig, S.P. Niclou, and E. Gottlieb. 2015. Glutamine synthetase activity fuels nucleotide biosynthesis and supports growth of glutamine-restricted glioblastoma. Nature Cell Biology. 17: 1556–1568.

213. Altman, B.J., Z.E. Stine, and C.V. Dang. 2016. From Krebs to clinic: glutamine metabolism to cancer therapy. Nat. Rev. Cancer. 16: 619–634.

214. Jiang, L., A.A. Shestov, P. Swain, C. Yang, S.J. Parker, Q.A. Wang, L.S. Terada, N.D. Adams, M.T. McCabe, B. Pietrak, S. Schmidt, C.M. Metallo, B.P. Dranka, B.

Schwartz, and R.J. DeBerardinis. 2016. Reductive carboxylation supports redox homeostasis during anchorage-independent growth. Nature. 532: 255–258.

215. Gross, M.I., S.D. Demo, J.B. Dennison, L. Chen, T. Chernov-Rogan, B. Goyal, J.R. Janes, G.J. Laidig, E.R. Lewis, J. Li, A.L. Mackinnon, F. Parlati, M.L.M. Rodriguez, P.J. Shwonek, E.B. Sjogren, T.F. Stanton, T. Wang, J. Yang, F. Zhao, and M.K. Bennett. 2014. Antitumor activity of the glutaminase inhibitor CB-839 in triple-negative breast cancer. Mol. Cancer Ther. 13: 890–901.

216. Kung, H.-N., J.R. Marks, and J.-T. Chi. 2011. Glutamine Synthetase Is a Genetic Determinant of Cell Type–Specific Glutamine Independence in Breast Epithelia. PLoS Genet. 7: e1002229.

217. Jain, M., R. Nilsson, S. Sharma, N. Madhusudhan, T. Kitami, A.L. Souza, R. Kafri, M.W. Kirschner, C.B. Clish, and V.K. Mootha. 2012. Metabolite Profiling Identifies a Key Role for Glycine in Rapid Cancer Cell Proliferation. Science. 336: 1040–1044.

218. Amelio, I., F. Cutruzzolá, A. Antonov, M. Agostini, and G. Melino. 2014. Serine and glycine metabolism in cancer. Trends in Biochemical Sciences. 39: 191–198.

219. Mattaini, K.R., M.R. Sullivan, and M.G. Vander Heiden. 2016. The importance of serine metabolism in cancer. J. Cell Biol. 214: 249–257.

220. Jerby, L., L. Wolf, C. Denkert, G.Y. Stein, M. Hilvo, M. Oresic, T. Geiger, and E. Ruppin. 2012. Metabolic Associations of Reduced Proliferation and Oxidative Stress in Advanced Breast Cancer. Cancer Research. 72: 5712–5720.

221. Possemato, R., K.M. Marks, Y.D. Shaul, M.E. Pacold, D. Kim, K. Birsoy, S. Sethumadhavan, H.-K. Woo, H.G. Jang, A.K. Jha, W.W. Chen, F.G. Barrett, N. Stransky, Z.-Y. Tsun, G.S. Cowley, J. Barretina, N.Y. Kalaany, P.P. Hsu, K. Ottina, A.M. Chan, B. Yuan, L.A. Garraway, D.E. Root, M. Mino-Kenudson, E.F. Brachtel, E.M. Driggers, and D.M. Sabatini. 2011. Functional genomics reveal that the serine synthesis pathway is essential in breast cancer. Nature. 476: 346–U119.

222. Delage, B., D.A. Fennell, L. Nicholson, I. McNeish, N.R. Lemoine, T. Crook, and P.W. Szlosarek. 2010. Arginine deprivation and argininosuccinate synthetase expression in the treatment of cancer. International Journal of Cancer. 13: NA–NA.

223. Husson, A., C. Brasse-Lagnel, A. Fairand, S. Renouf, and A. Lavoinne. 2003. Argininosuccinate synthetase from the urea cycle to the citrulline-NO cycle. Eur J Biochem. 270: 1887–1899.

224. Lind, D.S. 2004. Arginine and cancer. J. Nutr. 134: 2837S–2841S– discussion 2853S.

225. WU, G., and S.M. MORRIS Jr. 1998. Arginine metabolism: nitric oxide and beyond. Biochemical Journal. 336: 1–17.

226. Gerner, E.W., and F.L. Meyskens. 2004. Polyamines and cancer: old molecules, new understanding. Nat. Rev. Cancer. 4: 781–792.

227. Caso, G., M.A. McNurlan, N.D. McMillan, O. Eremin, and P.J. Garlick. 2004. Tumour cell growth in culture: dependence on arginine. Clin. Sci. 107: 371–379.

228. Gonzalez, G.G., and C.V. Byus. 1991. Effect of dietary arginine restriction upon ornithine and polyamine metabolism during two-stage epidermal carcinogenesis in the mouse. Cancer Research. 51: 2932–2939.

229. Wheatley, D.N., and E. Campbell. 2003. Arginine deprivation, growth inhibition and tumour cell death: 3. Deficient utilisation of citrulline by malignant cells. Br. J. Cancer. 89: 573–576.

230. Rabinovich, S., L. Adler, K. Yizhak, A. Sarver, A. Silberman, S. Agron, N. Stettner, Q. Sun, A. Brandis, D. Helbling, S. Korman, S. Itzkovitz, D. Dimmock, I. Ulitsky, S.C.S. Nagamani, E. Ruppin, and A. Erez. 2015. Diversion of aspartate in ASS1-deficient tumours fosters de novo pyrimidine synthesis. Nature. 527: 379–383.

231. Long, Y., W.-B. Tsai, D. Wang, D.H. Hawke, N. Savaraj, L.G. Feun, M.-C. Hung, H.H.W. Chen, and M.T. Kuo. 2017. Argininosuccinate synthetase 1 (ASS1) is a common metabolic marker of chemosensitivity for targeted arginine- and glutamine-starvation therapy. Cancer Lett. 388: 54–63.

232. Scott, L., J. Lamb, S. Smith, and D.N. Wheatley. 2000. Single amino acid (arginine) deprivation: rapid and selective death of cultured transformed and malignant cells. Br. J. Cancer. 83: 800–810.

233. Takaku, H., M. Takase, S. Abe, H. Hayashi, and K. Miyazaki. 1992. In vivo anti-tumor activity of arginine deiminase purified from Mycoplasma arginini. International Journal of Cancer. 51: 244–249.

234. Takaku, H., M. Matsumoto, S. Misawa, and K. Miyazaki. 1995. Anti-tumor activity of arginine deiminase from Mycoplasma argini and its growth-inhibitory mechanism. Jpn. J. Cancer Res. 86: 840–846.

235. Swinnen, J.V., K. Brusselmans, and G. Verhoeven. 2006. Increased lipogenesis in cancer cells: new players, novel targets. Curr Opin Clin Nutr Metab Care. 9: 358–365.

236. Beckers, A., S. Organe, L. Timmermans, K. Scheys, A. Peeters, K. Brusselmans, G. Verhoeven, and J.V. Swinnen. 2007. Chemical inhibition of acetyl-CoA carboxylase induces growth arrest and cytotoxicity selectively in cancer cells. Cancer Research. 67: 8180–8187.

237. Zhan, Y., N. Ginanni, M.R. Tota, M. Wu, N.W. Bays, V.M. Richon, N.E. Kohl, E.S. Bachman, P.R. Strack, and S. Krauss. 2008. Control of cell growth and survival by enzymes of the fatty acid synthesis pathway in HCT-116 colon cancer cells. Clinical Cancer Research. 14: 5735–5742.

238. Bauer, D.E., G. Hatzivassiliou, F. Zhao, C. Andreadis, and C.B. Thompson. 2005. ATP citrate lyase is an important component of cell growth and transformation. Oncogene. 24: 6314–6322.

239. Hatzivassiliou, G., F. Zhao, D.E. Bauer, C. Andreadis, A.N. Shaw, D. Dhanak, S.R. Hingorani, D.A. Tuveson, and C.B. Thompson. 2005. ATP citrate lyase inhibition can suppress tumor cell growth. Cancer Cell. 8: 311–321.

240. Hart, T., M. Chandrashekhar, M. Aregger, Z. Steinhart, K.R. Brown, G. MacLeod, M. Mis, M. Zimmermann, A. Fradet-Turcotte, S. Sun, P. Mero, P. Dirks, S. Sidhu, F.P. Roth, O.S. Rissland, D. Durocher, S. Angers, and J. Moffat. 2015. High-Resolution CRISPR Screens Reveal Fitness Genes and Genotype-Specific Cancer Liabilities. Cell. 163: 1515–1526.

241. Blomen, V.A., P. Májek, L.T. Jae, J.W. Bigenzahn, J. Nieuwenhuis, J. Staring, R. Sacco, F.R. van Diemen, N. Olk, A. Stukalov, C. Marceau, H. Janssen, J.E. Carette, K.L. Bennett, J. Colinge, G. Superti-Furga, and T.R. Brummelkamp. 2015. Gene essentiality and synthetic lethality in haploid human cells. Science. 350: 1092–1096.

242. Clendening, J.W., and L.Z. Penn. 2012. Targeting tumor cell metabolism with statins. Oncogene. 31: 4967–4978.

243. Goard, C.A., M. Chan-Seng-Yue, P.J. Mullen, A.D. Quiroga, A.R. Wasylishen, J.W. Clendening, D.H.S. Sendorek, S. Haider, R. Lehner, P.C. Boutros, and L.Z. Penn. 2014. Identifying molecular features that distinguish fluvastatin-sensitive breast tumor cells. Breast Cancer Res. Treat. 143: 301–312.

244. Keyomarsi, K., L. Sandoval, V. Band, and A.B. Pardee. 1991. Synchronization of tumor and normal cells from G1 to multiple cell cycles by lovastatin. Cancer Research. 51: 3602–3609.

245. Dimitroulakos, J., W.H. Marhin, J. Tokunaga, J. Irish, P. Gullane, L.Z. Penn, and S. Kamel-Reid. 2002. Microarray and biochemical analysis of lovastatin-induced apoptosis of squamous cell carcinomas. Neoplasia. 4: 337–346.

246. Dimitroulakos, J., D. Nohynek, K.L. Backway, D.W. Hedley, H. Yeger, M.H. Freedman, M.D. Minden, and L.Z. Penn. 1999. Increased sensitivity of acute myeloid leukemias to lovastatin-induced apoptosis: A potential therapeutic approach. Blood. 93: 1308–1318.

247. Panieri, E., and M.M. Santoro. 2016. ROS homeostasis and metabolism: a dangerous liason in cancer cells. Cell Death Dis. 7: e2253.

248. Clendening, J.W., A. Pandyra, P.C. Boutros, S. El Ghamrasni, F. Khosravi, G.A. Trentin, A. Martirosyan, A. Hakem, R. Hakem, I. Jurisica, and L.Z. Penn. 2010. Dysregulation of the mevalonate pathway promotes transformation. Proc. Natl. Acad. Sci. U.S.A. 107: 15051–15056.

249. Ko, Y.J., and S.P. Balk. 2004. Targeting steroid hormone receptor pathways in the treatment of hormone dependent cancers. Curr Pharm Biotechnol. 5: 459–470.

250. Kremer, J.C., and B.A. Van Tine. 2017. Therapeutic arginine starvation in ASS1-deficient cancers inhibits the Warburg effect. Molecular & Cellular Oncology. 4: e1295131.

251. SCHIMKE, R.T. 1964. Enzymes of arginine metabolism in mammalian cell culture. I. Repression of argininosuccinate synthetase and argininosuccinase. Journal of Biological Chemistry. 239: 136–145.

252. Choudhari, S.K., M. Chaudhary, S. Bagde, A.R. Gadbail, and V. Joshi. 2013. Nitric oxide and cancer: a review. World J Surg Oncol. 11: 118.

253. Nakajima, E.C., and B. Van Houten. 2013. Metabolic symbiosis in cancer: refocusing the Warburg lens. Mol. Carcinog. 52: 329–337.

254. Sobol, I.M. 2001. Global sensitivity indices for nonlinear mathematical models and their Monte Carlo estimates.

255. Baracchini, E., and H. Bremer. 1987. Determination of synthesis rate and lifetime of bacterial mRNAs. Anal. Biochem. 167: 245–260.

256. Sengupta, N., S.T. Rose, and J.A. Morgan. 2011. Metabolic flux analysis of CHO cell metabolism in the late non-growth phase. Biotechnol. Bioeng. 108: 82–92.

257. Dietmair, S., M.P. Hodson, L.-E. Quek, N.E. Timmins, P. Chrysanthopoulos, S.S. Jacob, P. Gray, and L.K. Nielsen. 2012. Metabolite profiling of CHO cells with different growth characteristics. Biotechnol. Bioeng. 109: 1404–1414.

258. Altamirano, C., A. Illanes, S. Becerra, J.J. Cairó, and F. Gòdia. 2006. Considerations on the lactate consumption by CHO cells in the presence of galactose. J. Biotechnol. 125: 547–556.

259. Schuetz, R., L. Kuepfer, and U. Sauer. 2007. Systematic evaluation of objective functions for predicting intracellular fluxes in Escherichia coli. Mol Syst Biol. 3: 119.

260. Henry, O., M. Jolicoeur, and A. Kamen. 2011. Unraveling the metabolism of HEK-293 cells using lactate isotopomer analysis. Bioprocess Biosyst Eng. 34: 263–273.

261. Dietmair, S., M.P. Hodson, L.-E. Quek, N.E. Timmins, P. Gray, and L.K. Nielsen. 2012. A Multi-Omics Analysis of Recombinant Protein Production in Hek293 Cells. PLoS ONE. 7: e43394.

262. Qin, X.-Y., F. Wei, M. Tanokura, N. Ishibashi, M. Shimizu, H. Moriwaki, and S. Kojima. 2013. The effect of acyclic retinoid on the metabolomic profiles of hepatocytes and hepatocellular carcinoma cells. PLoS ONE. 8: e82860.

263. Munger, J., B.D. Bennett, A. Parikh, X.-J. Feng, J. McArdle, H.A. Rabitz, T. Shenk, and J.D. Rabinowitz. 2008. Systems-level metabolic flux profiling identifies fatty acid synthesis as a target for antiviral therapy. Nat Biotechnol. 26: 1179–1186.

264. Smalley, M., K. Leiper, D. Floyd, M. Mobberley, T. Ryder, C. Selden, E.A. Roberts, and H. Hodgson. 1999. Behavior of a cell line derived from normal human hepatocytes on non-physiological and physiological-type substrates: evidence for enhancement of secretion of liver-specific proteins by a three-dimensional growth pattern. In Vitro Cell. Dev. Biol. Anim. 35: 22–32.

265. Hirayama, A., K. Kami, M. Sugimoto, M. Sugawara, N. Toki, H. Onozuka, T. Kinoshita, N. Saito, A. Ochiai, M. Tomita, H. Esumi, and T. Soga. 2009. Quantitative metabolome profiling of colon and stomach cancer microenvironment by capillary electrophoresis time-of-flight mass spectrometry. Cancer Research. 69: 4918–4925.

266. Kami, K., T. Fujimori, H. Sato, M. Sato, H. Yamamoto, Y. Ohashi, N. Sugiyama, Y. Ishihama, H. Onozuka, A. Ochiai, H. Esumi, T. Soga, and M. Tomita. 2013. Metabolomic profiling of lung and prostate tumor tissues by capillary electrophoresis time-of-flight mass spectrometry. Metabolomics. 9: 444–453.

267. Ramírez, G., J.S. Hagood, Y. Sanders, R. Ramírez, C. Becerril, L. Segura, L. Barrera, M. Selman, and A. Pardo. 2011. Absence of Thy-1 results in TGF-β induced MMP-9 expression and confers a profibrotic phenotype to human lung fibroblasts. Laboratory Investigation. 91: 1206–1218.

268. Ramos, C., M. Montaño, J. García-Alvarez, V. Ruiz, B.D. Uhal, M. Selman, and A. Pardo. 2001. Fibroblasts from idiopathic pulmonary fibrosis and normal lungs differ in growth rate, apoptosis, and tissue inhibitor of metalloproteinases expression. Am. J. Respir. Cell Mol. Biol. 24: 591–598.

269. Metallo, C.M., P.A. Gameiro, E.L. Bell, K.R. Mattaini, J. Yang, K. Hiller, C.M. Jewell, Z.R. Johnson, D.J. Irvine, L. Guarente, J.K. Kelleher, M.G. Vander Heiden, O. Iliopoulos, and G. Stephanopoulos. 2011. Reductive glutamine metabolism by IDH1 mediates lipogenesis under hypoxia. Nature. 481: 380–384.

270. Erro, E., J. Bundy, I. Massie, S.-A. Chalmers, A. Gautier, S. Gerontas, M. Hoare, P. Sharratt, S. Choudhury, M. Lubowiecki, I. Llewellyn, C. Legallais, B. Fuller, H. Hodgson, and C. Selden. 2013. Bioengineering the liver: scale-up and cool chain delivery of the liver cell biomass for clinical targeting in a bioartificial liver support system. Biores Open Access. 2: 1–11.

271. Su, C., Z. Hou, C. Zhang, Z. Tian, and J. Zhang. 2011. Ectopic expression of microRNA-155 enhances innate antiviral immunity against HBV infection in human hepatoma cells. Virol. J. 8: 354.

272. Hofmann, U., K. Maier, A. Niebel, G. Vacun, M. Reuss, and K. Mauch. 2008. Identification of metabolic fluxes in hepatic cells from transient 13C-labeling experiments: Part I. Experimental observations. Biotechnol. Bioeng. 100: 344–

354.

273. Murphy, T.A., C.V. Dang, and J.D. Young. 2013. Isotopically nonstationary 13C flux analysis of Myc-induced metabolic reprogramming in B-cells. Metabolic Engineering. 15: 206–217.

274. Giskeødegård, G.F., H. Bertilsson, K.M. Selnæs, A.J. Wright, T.F. Bathen, T. Viset, J. Halgunset, A. Angelsen, I.S. Gribbestad, and M.-B. Tessem. 2013. Spermine and citrate as metabolic biomarkers for assessing prostate cancer aggressiveness. PLoS ONE. 8: e62375.

Curriculum Vitae

Joana Raquel PINTO VIEIRA

École Polytechnique Fédérale de Lausanne
EPFL / SB / ISIC / LCSB,
CH H4 594 (Bat. CH), Station 6,
CH-1015 Lausanne, Switzerland
Tel.: +41 21 69 37644
E-mail: joana.pintovieira@epfl.ch

EDUCATION

2017 **École Polytechnique Fédérale de Lausanne (EPFL), Switzerland**
Post-Doc / Scientific collaborator

2012-2017 **École Polytechnique Fédérale de Lausanne (EPFL), Switzerland**
PhD student / Research Assistant

2008-2010 **École Polytechnique Fédérale de Lausanne (EPFL), Switzerland**
MSc in Physics and Minor in Biomedical Engineering

2004-2008 **Faculdade de Ciências da Universidade do Porto, Portugal**
(Faculty of Sciences at the University of Porto)
"Licenciatura" in Astronomy (Physics and Applied Mathematics)

RESEARCH EXPERIENCE

2016 **Internship at Merrimack Pharmaceuticals, INC, Cambridge, MA, USA (12 months)**

Computational Modeler at Discovery Department

Project 1: Modeling mAb induced immune response

Project 2: Optimization of mAb titer in bioreactor
Project 3: Modeling of combinations of cancer treatment therapies

2012-2017 **École Polytechnique Fédérale de Lausanne (EPFL), Switzerland**
PhD studies. Supervisor: Prof. Vassily Hatzimanikatis
Topics:
- Mathematical modeling of mRNA translation in cell context for the identification of its rate-limiting steps.
- Study of the Warburg metabolic switch in mammalian cells.

2010-2011 **University of Bremen, Germany**
Research and Teaching Assistant. Supervisor: Prof. Klaus Pawelzik
Topic: Modeling of learning with a reward modulated spike-timing-dependent plasticity

2010 **École Polytechnique Fédérale de Lausanne (EPFL), Switzerland**
Master thesis. Supervisor: Prof. Paolo De Los Rios
Topic: Markov chain Monte Carlo simulation for polymer adsorption on a flat surface.

2009 **École Polytechnique Fédérale de Lausanne (EPFL), Switzerland**
Semester project. Supervisor: Prof. Olaf Blanke
Modeling of the Rubber Hand Illusion: A Bayesian approach for multisensory integration.

2008-2009 **École Polytechnique Fédérale de Lausanne (EPFL), Switzerland**
Semester project. Supervisor: Prof. Giovanni Dietler
Topics: a) Measurements of properties of biomolecules using AFM: Measuring persistence length and end-to-end distance; b) Study of DNA melting using AFM.

TEACHING EXPERIENCE

École Polytechnique Fédérale de Lausanne (EPFL), Switzerland

Teaching Assistant for the following courses: Computational Biotechnology Lab; Principles and Applications of; Systems Biology; Numerical Methods in Chemistry; Chemistry TP II (Computational module); Physics IV

University of Bremen, Germany

Teaching Assistant for the following courses: Computer and Software; Theoretical Neuroscience

GRANTS

Awarded grant within the COST scientific program on European systems genetics network to attend the 6th International Summer School on Emerging Technologies in Biomedicine: Bioinformatics and Systems Biology Approaches for the Analysis of Complex Biological Networks, Patras, Greece, July, 2012.

LIST OF PUBLICATIONS

Journals

J. Vieira, M. Masid, A. Chiappino-Pepe, M. Ataman, and V. Hatzimanikatis (2017) Thermodynamics-based flux balance analysis of cancer physiology
Status: *In preparation (provisory title)*

J. Vieira, J. Racle, and V. Hatzimanikatis (2016) *Analysis of Translation Elongation Dynamics in the Context of an Escherichia coli Cell*. Biophysical Journal

Conferences (Selected talks)

J. R. Pinto Vieira, J. Racle and V. Hatzimanikatis. *Computational analysis of the interplay between tRNA competitive behavior and decoding type on the modulation of translation elongation rate.* ICSB 2013 - 14th International Conference on Systems Biology, Copenhagen, Denmark, August 30th - September 3rd, 2013.

Conferences (Posters)

J. R. Pinto Vieira, M. Masid, M. Ataman, V. Hatzimanikatis
Understanding reprogramming of cancer cell metabolism using metabolic modeling.
7th LIMNA Symposium, CHUV, Lausanne, Switzerland, April 4, 2017

J. R. Pinto Vieira, M. Ataman and V. Hatzimanikatis. *Characterization of intracellular metabolic states in the origin of the metabolic reprogramming of cancer cells.* Biochemical and Molecular Engineering XIX, Puerto Vallarta, Mexico, July 12-16, 2015.

J. R. Pinto Vieira, A. Kiparissides and V. Hatzimanikatis. *Characterization of intracellular metabolic states in cells presenting the Warburg phenotype*. Metabolic Engineering X Conference, Vancouver, BC, Canada, June 15-19, 2014.

J. R. Pinto Vieira, A. Kiparissides and V. Hatzimanikatis. *Characterisation of intracellular metabolic states in cells presenting the Warburg phenotype.* COBRA 2014 - 3rd Conference on Constraint-Based Reconstruction and Analysis, Wintergreen Resort, Charlottesville, VA, USA, May 20-23, 2014.

J. R. Pinto Vieira, J. Racle and V. Hatzimanikatis. *Analysis of the dynamics and competition for resources in the system of protein translation using stochastic simulations.* 9th European Biophysics Congress, Lisbon. Portugal, July 13-17, 2013.

Joana Vieira, Orlando Arévalo and Klaus Pawelzik. *Stochastic gradient ascent learning with spike timing dependent plasticity.* Cognitive Neuroscience Society Conference (CNS), Stockholm, Sweden, July, 2011.

LANGUAGE SKILLS

Portuguese (native), English (fluent), French (fluent)

OTHER INTERESTS

Acrylic painting, drawing, writing fiction and fitness.