Towards comprehensive and consistent kinetic models of metabolism under uncertainty

Thèse N° 9194

Présentée le 8 mars 2019

à la Faculté des sciences de base
Laboratoire de biotechnologie computationnelle des systèmes
Programme doctoral en chimie et génie chimique

pour l’obtention du grade de Docteur ès Sciences

par

Tuure Eelis HAMERI

Acceptée sur proposition du jury

Dr A.-S. Chauvin, présidente du jury
Prof. V. Hatzimanikatis, directeur de thèse
Prof. R. Gunawan, rapporteur
Dr L. Wu, rapporteur
Prof. D. Bonvin, rapporteur

2019

ÉCOLE POLYTECHNIQUE
FÉDÉRALE DE LAUSANNE

Acknowledgements

In the words of Steve Jobs, “You can't connect the dots looking forward; you can only connect them looking backwards. So you have to trust that the dots will somehow connect in your future.” As I am writing these acknowledgements, I realize that the accomplishment of this thesis would not have been possible without several people who helped me “connect the dots”.

I would firstly like to thank Prof. Vassily Hatzimanikatis, my supervisor and academic father, who helped me pave my path towards completing this unforgettable and stimulating journey. You were not only a supervisor; you were always there for any advice (scientific or non-scientific) and for sharing your knowledge from so many fields. Furthermore, your challenging and criticism of my work - although sometimes source of frustration - combined with the freedom and the trust you gave me, certainly pushed me to develop as a scientific researcher. Thank you for this wonderful and unique opportunity!

Besides my advisor, I would also like to thank my jury president Dr. Anne-Sophie Chauvin and my thesis committee experts: Prof. Dominique Bonvin, Prof. Rudiyanto Gunawan and Dr. Liang Wu. Your dedication, encouragements, insightful comments and stimulating questions made my thesis defense a very memorable milestone in my life.

To my collaborators from Université de Lausanne, Prof. Valérie Chavez and Dr. Marc-Olivier Boldi, thank you for your huge and contagious enthusiasm that helped me power through these final months of the thesis. We can now rigorously quantify how much we know (or don’t know?) from our model predictions.

To my colleagues, former and current, Meric, Alex, Tiziano, Georgios S., Daniel H., Stepan, Stefano, Julien, Noushin, Anush, Arti, Jasmin, Homa, Zhaleh, Maria, Christine, Daniel W., Vikash and Liliana, thank you all for your kindness, sharing your knowledge and helpful support. Particularly, I would like to acknowledge Georgios F. for being such a nice and calm academic big brother who was always there when requiring some guidance and peace of mind, Misko with his great wisdom topped up with some Balkan humor when going through challenging times, Pierre and Robin for teaching me GIT, reproducibility and … reproducibility, Sophia for our coffee breaks and the many epic nights out (particularly the

interview evening with Tiziano?), Joana for all the fitness/diet discussions and for our midnight chats when you were in Boston, Yves for the lunch breaks (next time with your cousin?) and, last but not least, Milenko for Rostock, Heidelberg, lunch and coffee breaks, your jokes and great support when needed. And, to all of you, I am very thankful for your friendship.

To all of my friends from Lausanne, the city of angels, or Laus'Angeles should I say, thank you for making this place my new home. Natael for the endless discussions about life, society and astrophysics and, for those very tight chess and poker matches. Palma for your very witty humor and some of the many prescribed nights we need to let stress out. Gleb, still can't believe the coincidence that we ended up living in the same building by chance after UCL times. Mike, it's been 20 years… To Guillaume, Constance, Eric, Alan, Etienne and Delphine, and any others that I may be forgetting, thank you for your friendship and for showing me some good times in Lausanne.

Tania, thank you for always trusting me and helping me to find my way, even in the most meandering parts of my thesis and personal life. I am out of words to describe how grateful I am to you. Thank you very much!

My deepest gratitude goes to my family, my parents Irma and Ari-Pekka and, my brother Ilkka. Thank you for your continuous love and for always being so supportive of my life decisions. You gave me the unshakable Finnish "sisu", that inner strength, to strive towards my goals during this journey called life. I trust that you understand how much I love you all.

Finally, I would like to profoundly thank, my biggest supporter and love, Isabelle, for putting up with the emotions I have been through during this final stretch of the thesis. Through those majestic ups, and even worse downs, our relationship kept me going and helped me to eventually "connect the dots" and wrap up this thesis. Thanks my love!

Lausanne, December 21st, 2018

Abstract

Metabolism is the sum of the chemical interactions occurring inside a cell that process nutrients into cellular constituents and energy. Research in the recent decades has enabled the characterization of metabolic reaction networks for multiple organisms. Genome-scale models (GEMs) have been utilized to mathematically describe and analyze these networks. GEMs are stoichiometric models that can serve as basis for studying metabolism and identifying metabolic states. However, stoichiometric models of metabolism do not account for the dynamics of the system. An understanding of cellular dynamics and regulation is essential for metabolic engineering of cells that produce certain metabolites of interest. Kinetic models can provide invaluable knowledge about system dynamics of cellular metabolism as metabolic control analysis (MCA) can offer information about the system's response to perturbations. Nonetheless, the construction of kinetic models is a challenging endeavor as several hurdles have to be overcome. Kinetic models of metabolism are subject to two general issues: diversity of network topologies and underlying uncertainty. Kinetic models are often built on an *ad hoc* basis without clear explanation or justification about their network contents, as there is no systematic protocol for their construction. Furthermore, kinetic modeling is subject to uncertainty from various sources. Multiple steady states can characterize an observed physiology. Additionally, the kinetic mechanisms describing a system and the values of their kinetic parameters are often not known.

In this thesis, we tackle several issues that hinder the formulation of kinetic models. Firstly, we apply model reduction algorithms to systematically reduce GEMs to construct study-specific kinetic models of different degrees of complexity. We demonstrate that the MCA outputs for these kinetic models are mostly independent of the complexity level because we preserve model equivalency. Secondly, as published kinetic models to date are constructed around a steady state of metabolism, we analyzed the impact of alternative steady states on metabolic engineering strategies. We proposed a systematic workflow for deriving conclusions that take into account the alternative feasible steady states of a physiology. We concluded that MCA outputs are more sensitive to the metabolite concentrations than to the metabolic fluxes of the system. Hence, it is essential to consider alternative metabolic states for a given physiology. Thirdly, since multiple kinetic models can describe a physiology due to the given uncertainties, it is important to consider them in populations. In order to derive conclusions from populations of kinetic models, it is essential to quantify with what certainty we make such predictions. We tested different statistical methods for assigning confidence levels to our conclusions and make recommendations applicable to any kinetic models. Finally, we implemented a sensitivity analysis approach that can elucidate which input parameters of a kinetic model contribute the most to the uncertainty in MCA outputs. The approach appears to predict correctly the sources of uncertainty and can be applied to any large-scale kinetic models. Overall, the work from this thesis contributes towards establishing a systematic workflow for building more consistent and comprehensible kinetic models for observed physiologies under uncertainty.

Keywords: metabolism, metabolic engineering, kinetic model, uncertainty, topology, model complexity, metabolic control analysis (MCA), steady state, multiplicity, confidence level, sensitivity analysis.

Résumé

Le métabolisme est l'ensemble des interactions chimiques qui se déroulent au sein d'une cellule pour transformer des nutriments en constituants cellulaires et en énergie. Les recherches menées au cours des dernières décennies ont permis de décrire des réseaux de réactions métaboliques pour plusieurs organismes, et des modèles à l'échelle du génome (GEM) ont été utilisés pour analyser ces réseaux et les décrire d'un point de vue mathématique. Les GEM sont des modèles stœchiométriques qui, s'ils servent de base à l'étude du métabolisme et à l'identification des états métaboliques, ne rendent pas compte de la dynamique d'un système. Or, une bonne compréhension de la dynamique et de la régulation cellulaires est indispensable à l'étude métabolique des cellules produisant des métabolites dignes d'intérêt. Les modèles cinétiques fournissent en revanche des informations précieuses sur la dynamique des systèmes en métabolisme cellulaire, étant donné que l'analyse du contrôle métabolique (MCA) met en lumière les réponses d'un système à des perturbations. Toutefois, la construction de modèles cinétiques est une démarche semée d'écueils. Les modèles cinétiques du métabolisme doivent faire face à deux obstacles: la diversité des topologies des réseaux et l'incertitude sous-jacente. Les modèles cinétiques sont souvent construits sur une base *ad hoc* sans pour autant qu'une explication claire ou qu'une quelconque justification ne soit fournie quant au contenu de leur réseau, et sans que leur construction ne soit systématiquement documentée. La modélisation cinétique est en outre soumise à différentes sources d'incertitude. Plusieurs états stationnaires sont susceptibles de caractériser la physiologie observée. De plus, les mécanismes cinétiques qui décrivent un système ainsi que les valeurs des paramètres cinétiques concernés demeurent souvent inconnues.

Dans cette thèse, nous abordons ces difficultés qui entravent la formulation de modèles cinétiques. D'abord, nous utilisons des algorithmes de réduction de modèles destinés à réduire systématiquement les GEM pour construire des modèles cinétiques spécifiques à l'étude envisagée et présentant des degrés de complexité différents. Nous démontrons que les résultats de la MCA pour ces modèles cinétiques sont principalement indépendants du niveau de complexité puisque l'équivalence est préservée. Deuxièmement, compte tenu du fait que les modèles cinétiques actuels sont généralement construits autour d'un état stationnaire du métabolisme, nous avons analysé son impact sur les conclusions en considérant les résultats de la MCA. Nous proposons un *workflow* systématique permettant de tirer des conclusions qui tiennent compte des états stationnaires possibles pour une physiologie donnée. Cela nous a permis d'en déduire que les résultats de la MCA sont plus sensibles aux concentrations de métabolites qu'aux flux métaboliques du système, ce qui signifie qu'il est essentiel de se pencher sur d'autres états métaboliques pour une même physiologie. Troisièmement, puisque plusieurs modèles cinétiques peuvent décrire une physiologie donnée en raison de la marge d'incertitude évoquée, il est important de les étudier en populations. Pour que des populations de modèles cinétiques donnent lieu à des conclusions, il convient de quantifier le degré de certitude de nos prédictions. Nous avons testé différentes méthodes statistiques pour attribuer un degré de confiance à nos conclusions et pour émettre des recommandations applicables à n'importe quel modèle cinétique. Enfin, nous avons développé une approche fondée sur l'analyse de sensibilité et capable de distinguer quels paramètres d'un modèle cinétique créent le plus d'incertitude dans les résultats de la MCA. Cette approche semble prédire correctement les sources d'incertitude et peut être appliquée à n'importe quel modèle cinétique à grande échelle. De manière générale, cette thèse pose les jalons d'un *workflow* systématique permettant de créer des modèles cinétiques plus cohérents et compréhensibles pour les physiologies observées en conditions d'incertitude.

Mots-clés: métabolisme, ingénierie métabolique, modèle cinétique, incertitude, topologie, complexité des modèles, analyse du contrôle métabolique (MCA), états stationnaires, multiplicité, degré de confiance, analyse de sensibilité.

Table of Contents

Acknowledgements 3

Abstract 5

Résumé 7

List of figures 13

List of tables 19

List of abbreviations 21

1. Background 23

1.1. Mathematical modeling in metabolic engineering 24

1.1.1. Constraint-based modeling 24

1.1.2. Kinetic modeling 26

1.2. Towards consistent large-scale kinetic models of metabolism 27

1.3. Thesis outline 32

1.4. Articles included in this thesis 32

2. The effects of model complexity and size on metabolic flux distribution and control ... 33

2.1. Introduction 33

2.2. Results and discussion 35

2.2.1. Reduced *E.coli* models 35

2.2.2. Thermodynamic-based variability analysis 38

2.2.3. Model equivalency 39

2.2.3.1. Scaling up steady-states 39

2.2.3.2. Equivalence in kinetic parameters 42

2.2.4. Consistency in MCA across models 42

2.2.4.1. Ranking enzymes for flux control 42

2.2.4.2. MCA consistency across reduced models 44

2.2.4.3. Study of uncertainty in MCA 48

2.3. Conclusion 50

2.4. Materials and methods 52

2.4.1. Model reduction 53
2.4.2. Flux directionality assumptions 54
2.4.3. Thermodynamic analysis 55
2.4.4. Maximum equivalency between steady-states 55
2.4.5. Constructing kinetic models 56
2.4.6. Control coefficient deviation index 57
2.5. Appendix A **58**
3. Kinetic models of metabolism that consider alternative steady-state solutions of intracellular fluxes and concentrations **61**
3.1. Introduction **61**
3.2. Results and discussion **63**
3.2.1. Multiplicity of flux directionality profiles 64
3.2.2. Comparative analysis of alternative flux directionality profiles 67
3.2.2.1. Reference steady states (RSSs) of FDPs 67
3.2.2.2. Analysis of control patterns 71
3.2.2.3. Ranking target enzymes for flux control 72
3.2.2.4. Study of uncertainty in FCCs 74
3.2.3. Impact of flux and concentration profiles 76
3.2.3.1. Flux uncertainty propagation to control 77
3.2.3.2. Concentration uncertainty propagation to control 78
3.3. Conclusion **80**
3.4. Materials and methods **81**
3.4.1. Reduced *E. coli* model 81
3.4.2. Identification of alternative flux directionality profiles 82
3.4.3. Computation of reference and extreme steady states for alternative FDPs 82
3.4.4. Analysis of alternative solutions between FDPs 83
3.4.4.1. Thermodynamic displacement analysis 83
3.4.4.2. Kinetic parameter sampling 85
3.4.4.3. General statistics on FCCs across FDPs 85
3.4.5. Characterizing the distribution of kinetic parameters 86
3.4.5.1. Beta distributions 86
3.4.5.2. Implying prior beta distributions for sampling 87

3.4.6. Analysis of alternative solutions within FDPs 87
3.4.7. Metabolic Engineering and Synthetic Biology Design 88
3.5. Appendix B 88

4. Statistical inference in ensemble modeling of cellular metabolism 89
4.1. Introduction 89
4.2. Results and discussion 92
4.2.1. Kinetic model derivation 92
4.2.2. Uncertainty in flux control coefficients 94
4.2.3. Confidence intervals 94
4.2.4. Case study: mean difference confidence intervals 97
4.3. Conclusion 100
4.4. Materials and methods 101
4.4.1. Model reduction 101
4.4.2. Kinetic parameter sampling 101
4.4.3. Simultaneous CIs for variable significance 102
4.4.3.1. Univariate and simultaneous CI 102
4.4.3.2. Bonferroni’s simultaneous confidence interval (BCI) 102
4.4.3.3. The exact normal (ENCI) 103
4.4.3.4. Bootstrapped simultaneous CI (BootCI) 103
4.4.4. Confidence intervals for comparison of cases 104
4.4.5. Sample size calculation 104
4.5. Appendix C 105

5. Global sensitivity analysis of control coefficients derived with metabolic control analysis 107
5.1. Introduction 107
5.2. Results and Discussions 109
5.2.1. Kinetic model 110
5.2.2. Uncertainty from the pentose phosphate pathway 111
5.2.2.1. Computation of sensitivity indices 112
5.2.2.2. Kinetic model tightening 114
5.2.2.3. Application and validation 115

5.2.3. Future opportunities and limitations 117
5.3. Conclusions 119
5.4. Materials and methods 119
5.4.1. MCA, sampling saturations 119
5.4.2. GSA, calculating sensitivity indices 121
5.5. Appendix D 122
6. Conclusions and future perspectives 125
6.1. Conclusions 125
6.2. Future perspectives 130
6.2.1. Personalized medicine 131
6.2.2. Targeted metabolic engineering strategies 132
6.2.3. Models combining metabolism with signaling 133
Bibliography 135

List of figures

Figure 2. 1. *E.coli* network diagram illustrating the core topologies studied for D1, D2 and D3. Edges represent the metabolic reactions and the nodes correspond to the metabolites. The reactions (edges) and metabolites (nodes) are coloured according to their pertinence to D1, D2 and D3, in blue, red and green, respectively. The reaction labels are coloured in black when the reaction is unidirectional. The bidirectional reactions' labels are coloured if the reaction can operate in both forward and backward directions in D1, D2 or D3 in blue, red and green, respectively. The reactions that are bidirectional in a smaller model were also bidirectional in the larger model(s), as it is an expansion around the same core. Diagram does not include all the reactions of the systems. **36**

Figure 2. 2. Metabolites connecting subsystems and the number of pairwise connections they achieve. Central carbon metabolites that change in connectivity (A) between D1 and D2, and (B) between D2 and D3. .... **38**

Figure 2. 3. Illustration and analysis of core metabolism connections to biomass building blocks via lumped reactions. The schematic representation (A) of the studied metabolic networks where the edges are reactions and the nodes are metabolites. The blue edges and nodes belong to the core network of D1. The red edges and nodes correspond to additions from the D2 expansion. The dashed arrows are lumped reactions of the system, connecting the core to the biomass, where: black represents lumped reactions that are present in all the models, blue is for lumped reactions existing only prior to the expansion in D1 and red is for lumped reactions existing only after network expansion in D2. Biomass building blocks (BBBs) are represented by the green ovals. The black solid arrows indicate fluxomics data that were integrated for optimally grown *E.coli* [61]. Venn diagrams highlight differences in the lumped reactions of D1, D2 and D3 in terms of (B) subnetworks and in terms of (C) reactions. **41**

Figure 2. 4. Cellular growth (μ) control across models. Pairwise illustration of the union of the top 7 enzymes across the models in terms of absolute control over cellular growth for (A) D1 versus D2, and (B) D2 versus D3. The whiskers give the upper and lower quartiles of the FCC populations and the bars give the means. **43**

Figure 2. 5. *E.coli* network diagram illustrating the deviation index (DI) of reactions when scaling up from D1 to D2. The reactions added by the modular redGEM expansion from D1 to D2 are given in blue and the 271 enzymatic reactions in common between D1 and D2 are color-coded based on the DI. The reactions were categorized into low DI (0-25 percentile), medium DI (25-75 percentile) and high DI (75-100 percentile), and are given in dark green, light green and red, respectively. Diagram does not include all the reactions of the systems. **45**

Figure 2. 6. *E.coli* network diagram illustrating the deviation index (DI) of reactions when scaling up from D2 to D3. The reactions added by the modular redGEM expansion from D2 to D3 are given in blue and the 307 enzymatic reactions in common between D2 and D3 are color-coded based on the DI. The reactions were categorized into low DI (0-25 percentile), medium DI (25-75 percentile) and high DI (75-100 percentile), and are given in dark green, light green and red, respectively. Diagram does not include all the reactions of the systems. **47**

Figure 2. 7. Comparison of flux and absolute deviations in FCCs for glycolytic reactions for (A) D1 versus D2, and (B) D2 versus D3. The absolute deviations computed per subsystem correspond to the sum of the absolute deviations in FCCs of reaction i with respect to all enzymes of the subsystem j. The reactions contained in a subsystem are as defined in the original GEM that was reduced [77]. The flux values shown did not deviate by more than 1% between pairs of models. **49**

Figure 2. 8. Comparison of flux and absolute deviations in FCCs for PPP reactions for (A) D1 versus D2, and (B) D2 versus D3. The absolute deviations computed per subsystem correspond to the sum of the absolute deviations in FCCs of reaction i with respect to all enzymes of subsystem j. The reactions contained in a subsystem are as defined in the original GEM that was reduced [77]. The flux values shown did not deviate by more than 1% between pairs of models. **50**

Figure 2. 9. Workflow for building consistent reduced kinetic models. The various steps are discussed with further detail within the chapter. **53**

Figure 2. 10. Derivation of the deviation index (DI) from the summation theorem. **58**

Figure 3. 1. Procedure for characterizing and analyzing multiplicities in metabolic networks. The procedure consists of several computational steps wherein the available data are integrated, the alternative solutions are identified, the populations of non-linear models are built, and the output variables are analyzed to make robust conclusions (for details see main text). **64**

Figure 3. 2. Multiple operational configurations for the same observed physiology of aerobically grown *E. coli*. (A) Representation of *E. coli* network. The fluxomics data that were integrated are indicated as uptake, secretion and growth rates. The bidirectional reactions are colored: phosphoglucose isomerase, (PGI, magenta) and fumarase, (FUM, red). (B) Representation of the four FDPs for the physiology under study. (C) Flux and thermodynamic displacement distributions of PGI and FUM reactions for each of the four generated FDPs. The boxplots show distributions for 5,000 samples. The central red line indicates the median, and the bottom and top edges of the box indicate the 25th and 75th percentiles, respectively. The whiskers correspond to approximately ± 2.7, which is the standard deviation, or 99.3% coverage if the data are normally distributed. Outliers are the points not covered by the range of the whiskers and are plotted individually using the '+' symbol. The black diamond is the RSS value. Full metabolite names are given in supplementary materials (Appendix B1). **66**

Figure 3. 3. Optimally grown aerobic *E. coli* metabolic network. Each of the 10 reactions labeled in red has an associated graph with the respective flux and thermodynamic displacement distributions for each FDP. The boxplots show distributions for 5,000 samples. The central red line indicates the median, and the bottom and top edges of the box indicate the 25th and 75th percentiles, respectively. The whiskers correspond to approximately ± 2.7, the standard deviation, or 99.3% coverage if the data are normally distributed. Outliers are the points not covered by the range of the whiskers and are plotted individually using the '+' symbol. The black diamond is the RSS value. Full enzyme and metabolite names are given in supplementary materials (Appendix B1)... **69**

Figure 3. 4. Cluster analysis of thermodynamic displacements across FDPs. Heat map showing thermodynamic displacements of the reactions across each FDP. Reactions differed the most in thermodynamic displacement across FDPs based on categorization of displacements (Methods). The rows represent the similarity between reactions and the columns represent the similarity between FDPs. The distances between the dendrograms were computed based on the Euclidean distance between the thermodynamic displacements, both column- and row-wise. Full enzyme names are given in supplementary materials (Appendix B1). **70**

Figure 3. 5. General statistics on FCCs across FDPs. Histogram displaying the fraction of reactions that have FCCs with a certain sign pattern across the FDPs. There are three main categories of the FCCs: (i) consistent among all FDPs, (ii) FDP specific, and (iii) two FDPs contradicting two other FDPs. The FCCs were averaged over the 50,000 samples for each FDP, and the ones selected for analysis had a mean absolute value larger than 0.1 (10% fold change). For example, to assist reading the figure, the column “FDP 1 specific” has two possible scenarios as it contains FCCs that are positive in FDP1 and negative in the other three FDPs as well as FCCs that are negative in FDP1 and positive in the three other FDPs. **72**

Figure 3. 6. Flux control across FDPs for cellular growth, μ. Illustration of the union of the top five enzymes across the FDPs in terms of absolute control over cellular growth. The whiskers correspond to the upper and lower quartiles of the 50,000 FCC populations, and the bars correspond to the means. Full enzyme names are given in supplementary materials (Appendix B1). **74**

Figure 3. 7. Principal component analysis (PCA) of FCCs of cellular growth across FDPs. PCA was carried out on the cellular growth FCCs of the 50,000 samples for each FDP. Only cellular growth FCCs with respect to non-transporter and non-exporter reactions were considered. The two principal components, namely PC1 and PC2, were plotted to study the variance in the FCC samples. The values in brackets correspond to the variance covered by the principal components (PCs). Full enzyme names are given in supplementary materials (Appendix B1). **76**

Figure 3. 8. Flux control patterns across extreme steady-state solutions. Illustration of the union between the top 10 enzymes across the (A) flux and the (B) concentration ESSs in terms of absolute control over glucose uptake. FCCs were sorted in decreasing order of absolute magnitude of the RSSs (reference). The enzyme names in black indicate that the FCCs were sign-consistent (in agreement) and red if they were sign-inconsistent (in opposition). Full enzyme names are given in supplementary materials (Appendix B1). **79**

Figure 4. 1. Schematic diagram of workflow carried out in the study. Information about key steps of the workflow. **91**

Figure 4. 2. *E.coli* network diagram illustrating the core topology studied. Diagram does not include all the reactions of the systems. Full reaction and metabolite names are given in the supplementary (Appendix B1). **93**

Figure 4. 3. Top control coefficients for glucose transport (GLCptspp). The diamonds indicate the mean of the FCCs in decreasing order of absolute mean. CIs were derived using (A) univariate t-test, (B) Bonferroni, (C) exact normal and (D) bootstrapping (see Materials and Methods). The lower and upper whiskers correspond to the CIs. The CI is blue if it contains zero and, red if it does not. **95**

Figure 4. 4. Top flux control coefficients of glucose uptake (GLCptspp) with confidence intervals determined by different statistical approaches. The top 10 FCCs based on absolute mean are reported with diamonds. The whiskers indicate the CIs for univariate t-test (magenta), Bonferroni (blue), exact normal (red) and bootstrapping (black). The reader is referred to the Materials and Methods for technical details on CI computation. **96**

Figure 4. 5. Case study: differences of means using bootstrapping. Comparison of the differences in means of 15 FCCs for GLCptspp across 4 cases using the bootstrapping method (see Materials and Methods). The whiskers indicate the CIs and the diamonds report the estimates of the differences in means. The tests were carried out globally on the 90 estimates even though we report each case comparison as a separate plot. **99**

Figure 5. 1. Global sensitivity analysis workflow for characterizing sources of variability in large-scale kinetic models. Diagram providing details of the various steps required for the characterization of parameters responsible for variance in kinetic models and their outputs. **110**

Figure 5. 2. *E.coli* network diagram illustrating the reactions of the kinetic model. The reactions indicated in red correspond to the PPP reactions whose uncertainty we studied using variance-based global sensitivity analysis. Diagram does not include all the reactions of the system. **111**

Figure 5. 3. Sobol sensitivity indices for flux control coefficient with respect to PPP enzymes' saturation levels. Sobol sensitivity indices for St (green) and Si (yellow) of PPP enzymes for

(A) glucose uptake control coefficient with respect to glucose-6-phosphate dehydrogenase 2 (G6PDH2r) and (B) pyruvate transporter control coefficient with respect to ribulose-phosphate 3-epimerase (RPE). Enzyme saturation levels were sampled uniformly between 0 and 1. The 200’000 samples were split into three groups, and the mean and the standard deviation were calculated based on these. The whiskers indicate the standard deviation and the bars report the mean. **113**

Figure 5. 4. Sobol sensitivity indices for flux control coefficient with respect to PPP enzymes’ saturation levels. Sobol sensitivity indices for St (green) and Si (yellow) of PPP enzymes for (A) glucose uptake control coefficient with respect to glucose-6-phosphate dehydrogenase 2 (G6PDH2r) and (B) pyruvate transporter control coefficient with respect to ribulose-phosphate 3-epimerase (RPE). Enzyme saturation levels were sampled uniformly between 0 and 1 for the PPP-reactions. The non-PPP reactions’ saturation ranges were anchored to 10% around mid saturation levels determined from the mean value taken over a previous population of kinetic models (further explanations in the main text). The 350’000 samples were split into three groups, and the mean and the standard deviation were calculated based on these. The whiskers indicate the standard deviation and the bars report the mean. **115**

Figure 5. 5. Sobol sensitivity indices for flux control coefficient with respect to PPP enzymes’ saturation levels. Sobol sensitivity indices for St (green) and Si (yellow) of PPP enzymes for ethanol transporter (ETOHtrpp) with respect to transketolase (TKT1). Enzyme saturation levels were sampled uniformly between 0 and 1 for the PPP-reactions. The non-PPP reactions’ saturation ranges were anchored to 10% around mid saturation levels determined from the mean value taken over a previous population of kinetic models (further explanations in the main text). The 350’000 samples were split into three groups, and the mean and the standard deviation were calculated based on these. The whiskers indicate the standard deviation and the bars report the mean. **116**

Figure 6. 1. Workflow for constructing consistent kinetic models of metabolism.......... **129**

List of tables

Table 3. 1. Mean of flux control coefficients of a population of models. **86**

Table 5. 1. PPP reactions and their corresponding kinetic model Km parameters. These reactions correspond to the groups of parameters for which we studied their uncertainty contribution to flux control coefficients. **112**

List of abbreviations

BBB	Biomass building block
BCI	Bonferroni confidence interval
BootCI	Bootstrapping confidence interval
CART	Classification and regression tree
CI	Confidence interval
DGSM	Derivative-based global sensitivity measure
DNA	Deoxyribonucleic acid
E.coli	*Escherichia coli*
EM	Ensemble modeling
ENCI	Exact normal confidence interval
ESS	Extreme steady state
FBA	Flux balance analysis
FCC	Flux control coefficient
FDP	Flux directionality profile
GEM	Genome-scale metabolic model
GMCA	Global sensitivity analysis on metabolic control analysis
GSA	Global sensitivity analysis
MCA	Metabolic control analysis
MFA	Metabolic flux analysis
MILP	Multiple integer linear programming
MSC	Metabolic sensitivity coefficients
ODE	Ordinary differential equation
ORACLE	Optimization and risk analysis of complex living entities
PCA	Principal component analysis
RSS	Representative steady state
TFA	Thermodynamic-based flux analysis
TFBA	Thermodynamics-based flux balance analysis
TVA	Thermodynamic-based variability analysis

1. Background

Metabolism is the set of the biochemical reactions that occur inside all the cells of a living organism to transform nutrients into energy and organic materials in order to sustain life. Living organisms can also be utilized to produce certain metabolites of interest. In fact, one of the oldest examples dates back several millenniums with the usage of yeast for the manufacturing of alcoholic beverages. In this light, recent advances in biotechnology are enabling the construction of cell factories that utilize - mainly bacteria, yeast and plants - to produce certain specific materials of interest. The construction of cell factories relies on metabolic engineering, which is the use of genome editing to modify the metabolism of living organisms. Hence, understanding how metabolism is influenced by environmental factors and genetics is indispensable for performing metabolic engineering. Moreover, information about the enzymes that are rate-limiting or that have the highest impact on the phenotype of interest is quintessential to achieve certain metabolic engineering objective.

Biotechnological developments have facilitated DNA sequencing for multiple organisms and allowed researchers to connect genotype to phenotype. Indeed, annotation of the genome with metabolic functions elucidates the interconnectivity of the metabolites involved in the biological system. A stoichiometric matrix can mathematically represent the identified biochemical reactions occurring inside a cell. These stoichiometric matrices can serve as invaluable scaffolds for constructing dynamic models of metabolism that shed light onto the rate-limiting processes within a cell by unraveling regulatory properties of cellular metabolism. Understanding the complex regulatory mechanism of metabolism within cells could open up new avenues to metabolic engineers in the design of cell factories. In this thesis, we explore how mathematical modeling can provide information about regulatory characteristics of metabolism in living organisms and thus help in formulating metabolic engineering strategies.

1.1. Mathematical modeling in metabolic engineering

1.1.1. Constraint-based modeling

As increasing numbers of stoichiometric representations of the biochemical reactions occurring inside living entities and high-dimensional experimental data sets are becoming more abundant, more sophisticated methods for analyzing these are required. Constraint-based modeling has surfaced as the predominant approach for studying genome-scale metabolic networks. These networks are based on annotated genomes and experimental literature to integrate the known cellular biochemistry into stoichiometric matrices [1, 2]. Constraint-based methods can incorporate additional data into these mathematical representations with various formulations [3]. The models that encapsulate available data from different sources can then be used for physiological studies of an organism.

Flux balance analysis (FBA) [4-6] has been used extensively to study stoichiometric models of metabolic networks. FBA uses the stoichiometric matrix that describes metabolic reactions and a set of linear programming constraints to predict fluxes across reactions throughout the network. Consequently, a FBA model is constrained by both the mass balance by reaction stoichiometry and the constraints that can be imposed on fluxes in the form of inequalities. In FBA, tasks that the network has to perform to represent the physiology of the system are defined as objective functions or constraints and can be defined in different ways [7]. There are no explicit definitions of these tasks for all cells and physiologies but generally, maximization of biomass production has been presumed as one main objective in normally developing organisms [4, 6]. FBA problems are solved using simple linear problem algorithms to obtain flux distributions. The flux distributions give the possible fluxes through the reactions of a physiology based on the mass balances and the inequalities that constitute an underdetermined system [8, 9]. With the increased availability of omics data, efforts in constraint-based modeling have been made to further incorporate different types of data into models and thus reduce the flux ranges.

Several methodologies for incorporating thermodynamic constraints into models have been proposed [10-15]. Ataman and Hatzimanikatis reviewed these alternative methods and concluded that thermodynamic-based flux analysis (TFA), also known as Thermodynamics-based flux balance analysis (TFBA), provides the most complete formulation that incorporates the least possible bias on reaction directionalities into systems [16]. TFA uses a mixed-integer linear programming (MILP) formulation to integrate information about the second law of thermodynamics by estimating Gibbs free energy of reaction [17, 18]. The TFA constraints can be used to compute concentration ranges of metabolites within the predefined bounds and to include experimental metabolite concentration measurements into the model [13]. As the TFA problem still remains underdetermined, there is a solution space of metabolic states that can be sampled to gather study-specific populations of thermodynamically feasible steady-states [8, 9].

Further efforts have been made to integrate various types of data into metabolic models in order to improve their predictive capability. Certain studies have successfully incorporated gene expression data into constraint-based models in order to relate metabolic flux with gene expression levels [19-21]. More recent studies are further integrating proteomics data into such models in order to link gene and protein expression directly to metabolic flux [22-24]. Another alternative to these approaches has been proposed to constrain flux so that it does not exceed allowable maximum defined by a reaction's enzymatic properties [25]. It should be noted that these approaches neglect the physicochemical constraints that laws of thermodynamics define. Hence, TFA remains the most comprehensive constrain-based method to date for making metabolic steady-state predictions for phenotype without violating thermodynamics or introducing excessive bias about reaction directionalities [16]. These constraint-based models rely on the quasi-steady state assumption, which assumes that there is no accumulation of metabolites in the system. Yet, with these approaches, it is difficult to quantify the extent to which system perturbations affect metabolic outputs [26].

Work has been carried out to utilize FBA for studying system perturbations/dynamics. An iterative FBA formalism [4] was initially developed by Varma and Palsson to predict changes in metabolic pathway utilization upon cell culture environment modifications and has been applied to further study diauxic growth in *E.coli* [27]. The FBA problem is solved in multiple specific time instants in order to model dynamic change in cellular state. This approach can successfully predict transitions in cellular metabolic states but it does not account for the time taken by regulatory changes that have to occur at genetic and enzymatic levels to facilitate these changes [4]. Several recent developments [28-30] integrate various types of omics data into iterative FBA formalisms to examine dynamic cellular behavior in genome-scale models. However, these methodologies have been studied and compared previously, and it appears that these approaches may not adequately account for complex dynamic cellular behavior [31]. Information about cellular kinetics is necessary for mechanistically describing the impact of perturbations in enzyme levels and/or in metabolite concentrations on the system. Understanding how a cell regulates itself and responds to perturbations in its environment is crucial for the development of biofuels, petrochemicals, pharmaceutical products and specialty chemicals of interest.

1.1.2. Kinetic modeling

In metabolic engineering, we frequently attempt to increase specific target productivity by finding rate liming steps of a system and/or developing *de novo* pathways. Such studies enable the design of cell factories that produce desired target chemicals in novel and more efficient manners [32, 33]. Hence, kinetic models are essential for understanding the underlying dynamics of biological systems and for providing guidance in designing cell factories. In recent years, constraint-based modeling has been supplemented with attempts to integrate kinetic information into models to analyze metabolism and its regulation more comprehensively. Models incorporating kinetics should be able to assess, in theory, how modifying the activities of molecular components affect the performance of the cell.

Kacser and Burns suggested to use control coefficients to quantitatively describe a metabolic network's response to imposed perturbations [34]. In this light, metabolic control analysis (MCA) has surfaced as a tool facilitating the study of cellular regulation [34, 35]. MCA quantifies the extent to which reactions of a metabolic network affect and control fluxes and metabolic concentrations at a steady state [36]. As MCA is a local sensitivity analysis tool, the technique can only be applied to estimate impacts of small perturbations to biological systems [37]. Nevertheless, MCA is a well established method and has been successfully applied in the field of metabolic engineering, synthetic biology and disease treatment [38-42]. Furthermore, MCA does not require the evaluation of dynamic non-linear kinetic model integrals, making it computationally tractable when having to consider multiple alternative kinetic models.

As kinetic parameter data and knowledge about cell regulatory mechanisms are still scarcely available, constructing large-scale kinetic models is a demanding endeavor [43]. The uncertainty in the regulatory mechanisms and in the values of kinetic parameters results in the existence of multiple viable kinetic models, describing the same observed physiology [33]. Ensemble modeling (EM) [44-47] and Optimization and Risk Analysis of Complex Living Entities (ORACLE) [48-53] propose workflows for sampling the kinetic parameter space and for constructing populations of large-scale kinetic models around a given steady-state. The EM approach uses experimental data to prune the populations of kinetic models in order to find a unique dynamic model that describe cellular behavior. ORACLE constructs populations of kinetic models, where nonlinear mechanistic rate laws describe reaction kinetics, and their algebraic system of equations is solved to compute model parameters. No bias is imposed as the entire populations of viable kinetic models are considered. Additionally, ORACLE readily computes the MCA sensitivity coefficients for populations of kinetic models. These properties and the efficient sampling of kinetic parameters make ORACLE more suitable for building large-scale and genome-scale kinetic models [54, 55]. Despite these considerable advances in frameworks for building kinetic models of metabolism, several issues such as bringing models to larger scale and, description and consideration of uncertainty are hindering the advancement of large-scale kinetic models.

1.2. Towards consistent large-scale kinetic models of metabolism

Metabolic network models found in literature vary significantly in terms of form and size depending on the purpose of different studies. Researchers use smaller *ad hoc* models that capture the essential biological features of the system and decrease computational costs. However, the trade-off between simplification and consistency remains a pertinent issue in computational modeling [33]. For instance, Teusink and colleagues built a reduced kinetic model of *Saccharomyces cerevisiae* but their initial model did not converge to the experimentally observed state. They added branches to the network in order to reach the evidenced steady state [56]. The paper suggests that kinetic models can be curated in order to mirror experimentally observed *in vivo* phenomena but the model reduction process follows an *ad hoc* approach. Some reduced models focus on certain pathways and sub-systems instead of capturing genome-scale model features. Chassagnole and co-workers produced a dynamic model of central carbon pathway for *Escherichia coli* and applied MCA principles for studies. They compared simulations with experimental observations, which gave credibility to their simulations [39]. However, like the model from Teusink and team, the reduction is *ad hoc* and does not include a biomass reaction to simulate growth requirements [39, 56]. While these models have provided useful insights, they were neither thermodynamically consistent, nor stoichiometrically balanced.

The largest published kinetic model to date is the *k-ecoli457* genome-scale *Escherichia coli* model that was developed by Khodayari and Maranas [45]. Their modeling approach removes reactions that do not carry any flux under the experimentally observed physiological conditions and they then use a simplified biomass equation to model cellular building blocks' biosynthesis. This strategy works for querying/predicting properties of the modeled physiology but, the approach for simplifying the cellular biosynthesis function is physiology-specific and cannot be directly applied to other observed physiologies. Several methods have been proposed for reducing the complexity of metabolic networks in a systematic way [57-59]. To circumvent issues arising from case-specific model simplifications and *ad hoc* model reductions, Ataman and colleagues have developed algorithms, redGEM and lumpGEM, that reduce genome scale models systematically [57, 58]. The methods ensure that the reduced stoichiometric model is consistent with its genome-scale counterpart in terms of flux profiles, metabolic concentrations, Gibb's free energy and the stoichiometry of biomass [57, 58]. This approach provides promising opportunities for reducing the complexity of genome-scale models in order to build kinetic models that are more tractable and context-specific.

But to what extent do model complexity reductions change our conclusions on the cell behavior? Palsson and Lee have studied model complexity and its impact on the predictions of their kinetic models [60]. Their study suggested that model complexity highly affects numerical values of the MCA-derived sensitivity coefficients. However, they did not apply a systematic model reduction method. Sensitivity coefficients can be used to compare how the topology of the model impacts the regulatory properties of the network. Hence, these MCA sensitivity coefficients can serve as quantifiable outputs of kinetic model for measuring how network complexity affects model predictions. In **Chapter 2**, a workflow for building consistently reduced kinetic models around a steady-state that characterizes the system is suggested. The workflow applies redGEM and lumpGEM algorithms to build study-specific stoichiometric models of different complexity levels [57, 58]. A procedure for scaling up the steady-states between these modular models of different complexity is proposed. Additions to the ORACLE workflow were made to construct populations of kinetic models around the steady-states of the reduced stoichiometric models, whilst ensuring parametric equivalency between the populations. The proposed workflow allowed the construction of populations of study-specific kinetic models that shed light into the effect of model complexity on MCA outputs.

A combination of model reduction algorithms, constraint-based methods and experimental data can be used to construct more comprehensive kinetic models of metabolism. Despite these important advances, biological systems remain highly underdetermined and there are numerous possible steady-states that could characterize a physiology. These systems are usually continuous and discrete at the same time, resulting in complex and high-dimensional solution spaces. The modeled physiology of optimally grown *E.coli* [61] could not be uniquely defined and multiple flux distributions describe the system as some reactions are bidirectional [62]. Bidirectional reactions are reversible reactions that can operate in both forward and backward directions when applying the thermodynamics-based constraints of the network and any other imposed constraints [14]. However, in a flux directionality profile (FDP), reactions can only operate in a unique direction. Consequently, models of metabolism can have numerous possible FDPs, resulting in multiple feasible steady states due to bidirectional reactions [63]. Nevertheless, kinetic models to date are generally constructed around a unique steady-state describing a given physiology [55].

How does the selection of a steady-state impact the conclusions drawn from kinetic models? As the impact of the steady-state on kinetic models and their outputs has not been studied previously, it was essential to inquire how this uncertainty in model variables propagates into conclusions drawn from kinetic models. **Chapter 3** develops a workflow that takes into account the multiplicity of feasible steady-states describing a given physiology when constructing kinetic models. The solution space for an observed physiology can be separated into different FDPs where directionality is assigned to each bidirectional reaction to get alternative convex solution spaces for the different feasible states of the metabolic network. Different FDPs were compared in order to understand their impact on kinetic model outputs. The effects of concentration and flux were decoupled to study how their uncertainty affects conclusions drawn from MCA outputs.

The populations of MCA outputs describe the average behavior of the system and can highlight enzymes that could be of interest for genetic manipulation. Previous studies that considered distributions of MCA outputs compared their mean behavior and accounted for the uncertainty with errors bars that represent the first and third quartiles of the populations [48, 64]. A statistical method that can be used for accounting for uncertainty is to consider confidence intervals that define ranges for observed variables that will contain their estimated values for a given probability. Confidence intervals are constructed with underlying assumptions about the nature of the data. Usually they assume that the modeled data follows a normal behavior and the confidence intervals are generated independently for each variable. Work has already been done in incorporating this approach into kinetic modeling framework [65]. However, populations of MCA outputs will not have a normal distribution and it is important to consider alternative methods for deriving confidence intervals for such data in order to derive statistically significant conclusions.

In order to quantify the uncertainty in model outputs, confidence intervals can be derived using different statistical procedures. In **Chapter 4**, alternative statistical approaches for ranking distributions of MCA outputs were considered. The models from **Chapter 3** were used to compare how effectively these statistical methods can be applied for deriving meaningful conclusions. Furthermore, researchers in systems biology often have to compare different sets of data that could for instance stem from models of different physiologies. Hence, these different statistical approaches were also used for making meaningful comparisons of distributions. The methodologies were benchmarked in order to suggest a workflow for deriving conclusions from populations of MCA outputs. Discussions around the advantages and disadvantages of these methods provide guidelines that will help selecting the appropriate statistical method for deriving statistically significant conclusions. These methods can be readily applied for different sets of data encountered in systems biology and are not only limited to applications with MCA outputs.

Uncertainty stemming from the variables (concentration and flux) of kinetic models was studied in **Chapter 3**. However, kinetic models and their conclusions are subject to parametric uncertainty as they contain numerous parameters for which the numerical values are not known. Gutenkunst *et al.* suggested that a majority of the parameters of a kinetic model are “sloppy” and that generally, only several parameters will be constrained to narrow regions [66]. The iSCHRUNK classification algorithm was developed by Andreozzi *et al.* to underpin such parametric patterns that they call “rules” in populations of kinetic models [49]. Another alternative to such machine learning approaches is to perform global sensitivity analysis (GSA) on kinetic parameters. However, previous applications of GSA to kinetic models of metabolism are limited to smaller models [67]. **Chapter 5** demonstrates how parameter uncertainty can be studied with GSA using a variance-based approach. GSA has been performed in studies of metabolism on significantly smaller dynamic models. As the model complexity increases, the computational expenses increase exponentially and it becomes challenging to carry out such studies [68]. A workflow that applies a variance-based approach, developed by Saltelli *et al.*, was proposed for unraveling the contribution of kinetic parameters to the variance of study-specific MCA outputs [69]. The workflow can be used to source uncertainty and guide the design of experiments that will reduce uncertainty in parameters. Knowledge about the uncertainty associated with kinetic parameters is particularly useful as the techniques for experimentally estimating various kinetic parameters develop [70].

1.3. Thesis outline

From a global perspective, **Chapters 2-5** contribute towards the objective of defining an overall workflow that will guide the construction of comprehensive and consistent kinetic models of metabolism under uncertainty. **Chapter 2** considers how model complexity should be assessed in order to obtain consistent and context-specific kinetic models. In **Chapter 3**, the reader is exposed to the importance of considering alternative steady-states of metabolism. **Chapter 4** provides statistical tools that can be used to construct confidence intervals on kinetic model data to derive statistically significant conclusions. Then, **Chapter 5** presents an approach that can be applied to reduce uncertainty in kinetic models and further assist in experimental design. Finally, **Chapter 6** discusses around the applications of the workflows developed in this thesis and the future opportunities in the research field.

The supplementary material for this thesis is organized chapter-wise and is available in an electronic format at the following link:
https://zenodo.org/record/2534074#.XDR6ZJNKhBw

1.4. Articles included in this thesis

Hameri T, Fengos G, Ataman M, Miskovic L, Hatzimanikatis V. Kinetic models of metabolism that consider alternative steady-state solutions of intracellular fluxes and concentrations. Metabolic Engineering. 2018 (in press).

Hameri T, Fengos G, Hatzimanikatis V. The effects of model complexity and size on metabolic flux distribution and control. 2018 (in preparation).

Hameri T, Boldi M-O, Hatzimanikatis V. Statistical inference in ensemble modeling of cellular metabolism. 2018 (in preparation).

Denhardt-Eriksson R, **Hameri T**, Hatzimanikatis V. Global sensitivity analysis of control coefficients derived with metabolic control analysis. 2018 (in preparation).

2. The effects of model complexity and size on metabolic flux distribution and control

2.1. Introduction

Kinetic models of cellular metabolism can provide comprehensive understanding on the dynamics of the cell and its response to environmental changes and perturbations. In depth understanding of cellular metabolism can allow metabolic engineers to tailor cells according to sought specifications and objectives. This could enable the design of cell factories where flux is directed towards the production of biofuels, pharmaceuticals or other specialty chemicals. To be useful though, a kinetic model should represent the dynamics of the cell accurately enough to provide the required study-specific knowledge [71]. To date, important strides towards building large- and genome-scale kinetic models of metabolism have been made [48, 50, 72, 73]. Despite the emergence of methodologies for building kinetic models, the research community knows that several challenges remain to be confronted.

With larger and better quality kinetic models, the mathematical representations become increasingly complex. Furthermore, the parameter sensitivities of systems biology models are in general "sloppy" [66]. Hatzimanikatis and coworkers noted that metabolic models are often built around certain central carbon pathways or, *ad-hoc* reduced models of genome-scale metabolic network models (GEMs) [57]. Such models do not account for the full information contained in the GEMs and, the *ad hoc* reduced models do not come with explicit explanations and justifications on how the model was reduced. Several studies have built kinetic models around *ad hoc* reduced models and computed Metabolic Sensitivity Coefficients (MSCs) for the system with metabolic control analysis (MCA) [48, 64, 71, 74, 75]. MSCs are desirable outputs of the kinetic models as they give insight into control patterns of the cell, assuming that the model is correct and accurate. However, Palsson and

Lee showed that network complexity significantly affected the numerical values and the interpretation of MSCs [76]. Their study showed that three different red cell metabolic models produced MSCs that have opposite signs. This suggested that the analysis of incomplete metabolic models could lead to misleading and inaccurate information.

Palsson and Lee acknowledge that their models were reduced in an *ad hoc* manner to analyze how significantly network complexity can affect the MSCs of kinetic models [76]. However, nowadays algorithms for reducing GEMs are starting to emerge [57-59]. The NetworkReducer algorithm aims to reduce the network around certain “protected” metabolites and reactions by iteratively removing reactions that do not obstruct their activity [59]. Ataman and coworkers developed the redGEM and lumpGEM algorithms which allow reduction of GEMs around selected subsystems by retaining linkages and the information captured in GEMs [57, 58]. The algorithm performs consistency checks with the GEM to ensure that the reduced model is consistent in terms of flux profiles, essential genes and reactions, thermodynamic feasible ranges of metabolite concentrations and ranges of Gibbs free energy of reactions. The redGEM and lumpGEM algorithms can be used to build thermodynamically feasible models with different levels of complexity consistent with the GEM for the same chosen subsystems. These algorithms open up the possibility to investigate how MSCs are affected by model complexity for consistently reduced models by building kinetic models around them.

In this chapter, we used the redGEM and lumpGEM algorithm to reduce the *E.coli* iJO1366 GEM to three different models, namely D1, D2 and D3, encompassing 271, 307 and 327 enzymatic reactions and 160, 188 and 197 metabolites, respectively. The thermodynamic formulation of the stoichiometric models allowed integration of fluxomics and metabolomics data for aerobically grown *E.coli* (Appendix A1) [61]. Due to the topological differences between the three models, we proposed a technique for scaling up the flux profile and concentration vector reference steady states from D1 into the larger models D2 and D3. This scale-up procedure ensures physiological equivalency of the models by assuring that their steady states are numerically similar. All the three models satisfy thermodynamic constraints and are consistent with the GEM. We used the Optimization and Risk Analysis of Complex Living Entities (ORACLE) workflow to construct kinetic models for D1, D2 and D3 around their scaled reference steady states. We fixed kinetic parameters from the smaller model into the larger one to further ensure equivalency of the models and hence a fair comparison. As integral part of the ORACLE workflow, we compute the MSCs for the stable kinetic models. We studied the MSCs across the three models and demonstrate that there is consistency amongst MSCs and that we can make metabolic engineering decisions, independent of model complexity.

2.2. Results and discussion

2.2.1. Reduced *E.coli* models

We applied redGEM and lumpGEM algorithms [57, 58] to systematically derive modular, reduced, *E.coli* stoichiometric models (Methods) from the iJO1366 GEM [77]. We selected glycolysis, pentose phosphate pathway (PPP), tricarboxylic acid (TCA) cycle, glyoxylate cycle, pyruvate metabolism and electron transport chain (ETC) as the subsystems (as defined in the iJO1366 GEM [77]) around which reduction was performed to different degrees of connectivity, similarly to Ataman *et al.* [57]. These subsystems contain the 12 essential biomass precursors defined by Neidhart *et al.* [78] and capture the central carbon metabolism of *E.coli*. Reduced stoichiometric models D1, D2 and D3 inter-connect the subsystems between each other with up to one, two and three reactions, respectively. Consequently, D1, D2 and D3 model cores generated via redGEM are constituted of 271, 307 and 327 enzymatic reactions and 160, 188 and 197 metabolites, respectively (Figure 2.1). The cores were connected to biomass production via lumped reactions, generated by the lumpGEM, to characterize the rest of the GEM (further discussion on lumped reactions around Figure 2.2 and 2.3 later in this section).

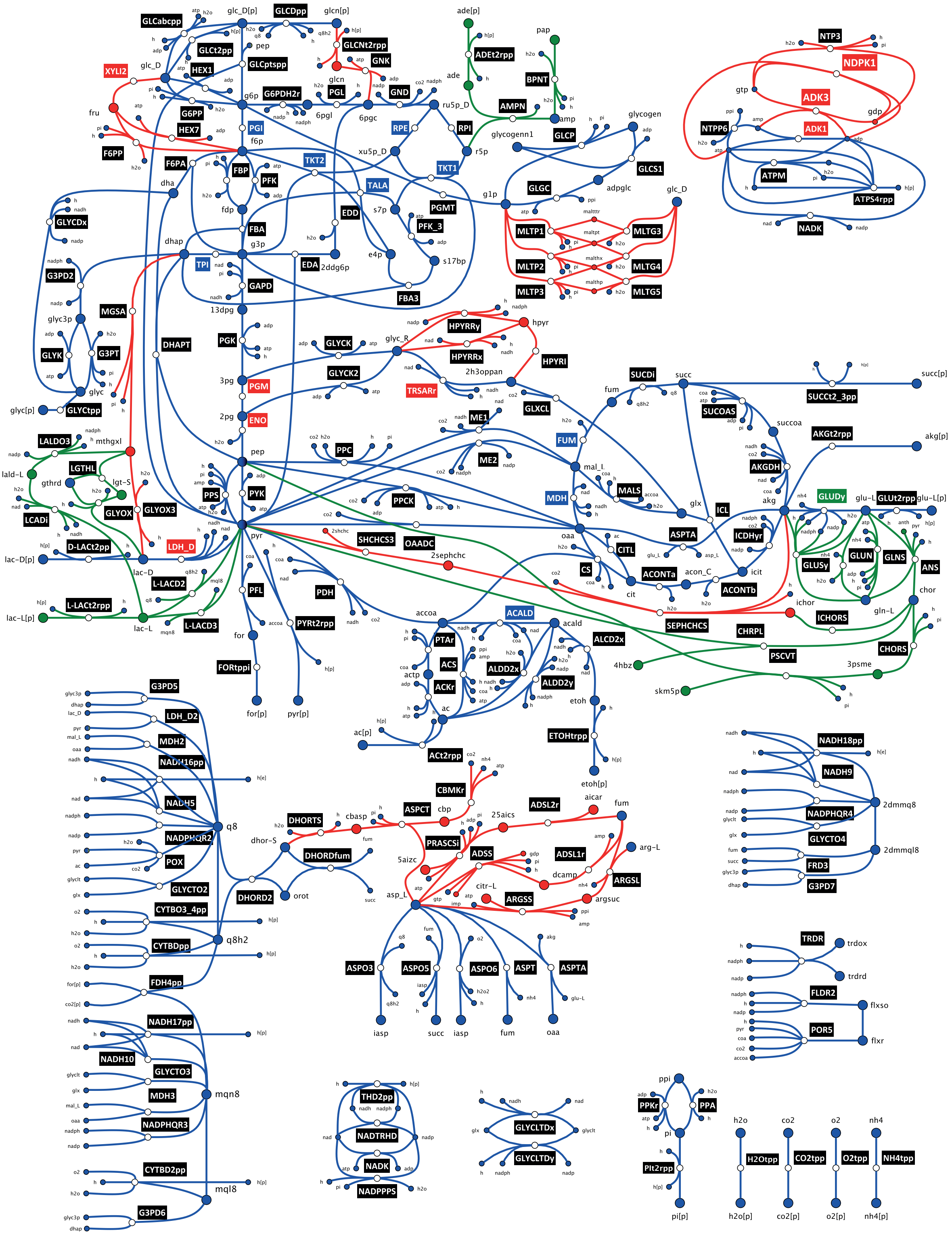


Figure 2. 1. ***E.coli*** **network diagram illustrating the core topologies studied for D1, D2 and D3. Edges represent the metabolic reactions and the nodes correspond to the metabolites. The reactions (edges) and metabolites (nodes) are coloured according to their pertinence to D1, D2 and D3, in blue, red and green, respectively. The reaction labels are coloured in black when the reaction is unidirectional. The bidirectional reactions' labels are coloured if the reaction can operate in both forward and backward directions in D1, D2 or D3 in blue, red and green, respectively. The reactions that are bidirectional in a smaller model were**

also bidirectional in the larger model(s), as it is an expansion around the same core. Diagram does not include all the reactions of the systems.

The additional reactions in D2 include xylose isomerase (XYLI2), hexokinase D-fructose (HEX7) and D-fructose 6-phosphate phosphatase (F6PP), that connect D-glucose with D-fructose 6-phosphate via D-fructose (Figures 2.1 and 2.2). D2 also includes the maltodextrin system which connects the D-glucose to D-glucose 1-phophate via the maltodextrin phosphorylase and maltodextrin glucosidase reactions. In D2, dihydroxyacetone phosphate can react to methyglyoxal, which in turn can react to D-Lactate, providing increased connectivity between glycolysis and the pyruvate node. Additionally, pyruvate can react to 2-succinyl-5-enolpyruvyl-6-hydroxy-3-cyclohexene-1-carboxylate, which can react to form 2-oxoglutarate, thus connecting the TCA cycle with the pyruvate metabolism. D2 also includes three different ways to connect with two reactions from fumarate to L-aspartate, which - via argininosuccinate, adenylsuccinate and adenylosuccinate - further link the TCA cycle with the ETC. The adentylate kinase (ADK3), nucleoside-diphosphate kinase (NDPK1) and nucleoside-triphosphatase (NTP3) enzymes provide D2 model with additional flexibility in the system's energy metabolism.

D3 has additional reactions enabling the transformation of methylglyoxal into D-lactate and L-lactate. Methylglyoxal is a hub metabolite that provides connectivity between upper glycolysis to the pyruvate node. The pyruvate and phosphoenolpyruvate nodes are connected to the TCA cycle via chorismate. Fruthermore, the glutamin and glutamate synthases provide additional flexibility in allowing conversion between L-glutamate and L-glutamine. In D3 the presence of AMP nucleosidase (AMPN) provides an additional connection between the PPP and the ETC. However, the expansion from D1 to D2 resulted in more central carbon metabolites that change in connectivity than the expansion from D2 to D3 (Figure 2.2). Several hub metabolites like methylglyoxal, isochorismate, pyruvate, D-lactate and 2-oxoglutarate change in connectivity between the three models.

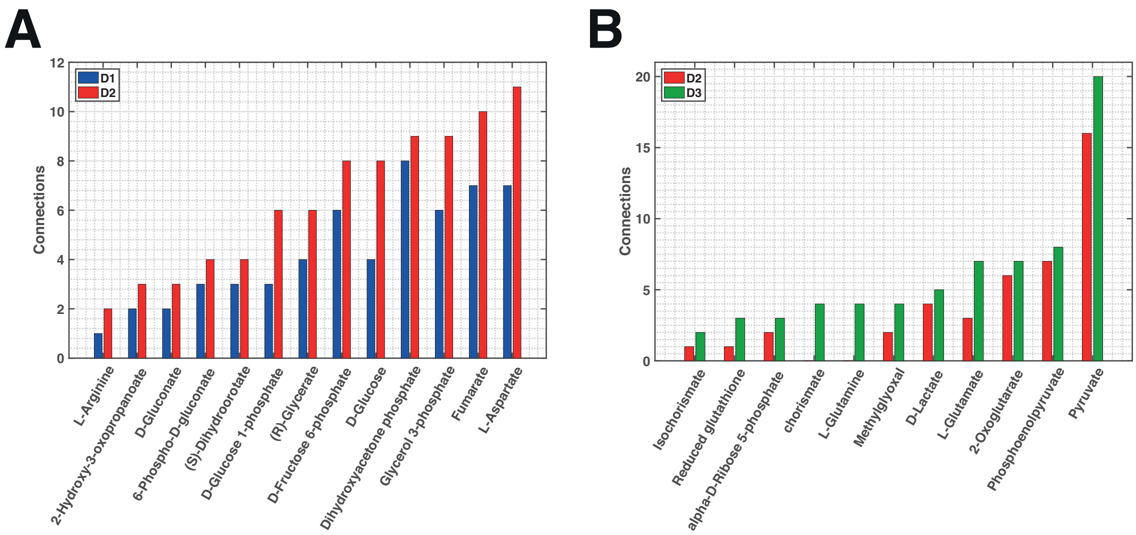


Figure 2. 2. Metabolites connecting subsystems and the number of pairwise connections they achieve. Central carbon metabolites that change in connectivity (A) between D1 and D2, and (B) between D2 and D3.

2.2.2. Thermodynamic-based variability analysis

Within the thermodynamic formulation [15] of the stoichiometric models D1, D2 and D3, we integrated fluxomics and metabolomics data for aerobically grown *E.coli* (Appendix A1). Several assumptions were made on reaction directionalities, based on literature [61, 79-82], to further constrain the models (Methods). We performed a thermodynamic-based variability analysis (TVA) [83] on D1, D2 and D3 and we found they had 9, 17 and 18 bi-directional reactions, respectively (Appendix A2).

Analysis of reaction flux ranges from TVA for D1, D2 and D3 revealed several considerable differences between the central carbon reactions. When comparing the ranges between D1 and D2, the largest differences were noted in adenylate kinase (ADK1), tartronate semialdehyde reductase (TRSARr), glucose-6-phosphate isomerase (PGI), phosphoglucomutase (PGMT), D-lactate dehydrogenase (LDH_D), triose-phosphate isomerase (TPI), phosphoglycerate kinase (PGK), glyceraldehyde-3-phosphate dehydrogenase (GAPD), phosphoglycerate mutase (PGM) and enolase (ENO) (Appendix A2). Performing the same analysis between D2 and D3 reveals that only transketolase (TKT2) and glutamate dehydrogenase (GLUDy) changed considerably in ranges amongst central carbon reactions. In fact, GLUDy became bidirectional with this expansion. Other differences were noted in the peripheral reactions pertaining to additions from expansion from D2 to D3. Generally, most differences in the flux variability ranges resulted from the bypass routes that additional reactions provided, which resulted in certain reactions becoming bi-directional (Figure 2.1).

The TVA was also performed on concentration ranges of the models and we notice several differences in the allowable ranges of metabolite concentrations between D1, D2 and D3. Most noticeable concentration range differences between D1 and D2 occurred in D-glucose 6-phosphate, D-fructose 6-phosphate, fumarate, L-arinine and S-Dihydroorotate (Appendix A2). However, between D2 and D3, noticeable differences were only noted in the ranges of D-glucose 6-phosphate and D-fructose 6-phosphate. The comparison of these TVA ranges in metabolite concentrations and reaction fluxes reveals that more considerable differences occurred between D1 and D2 than between D2 and D3.

2.2.3. Model equivalency

Despite the inclusion of omics data for aerobically grown *E.coli*, D1, D2 and D3 remained underdetermined systems, resulting in the existence of multiple alternative steady-states that can characterize the studied *E.coli* physiology. A representative steady-state is required for the flux profile and for the metabolite concentration vector, to build a kinetic model around the selected steady state. Furthermore, in light of benchmarking the outputs of kinetic models, the models are required to themselves be as close to each other as possible to allow for an unbiased comparison. Hence, their representative steady-states were kept similar so that the models describe the same operational state of the cell.

2.2.3.1. Scaling up steady-states

We sampled the flux and the concentration solution spaces for D1 and we used PCA to select representative steady states (Methods). To preserve equivalency across the kinetic models, it was desirable that the flux profile and the concentration vector steady states in D2 and D3 resemble the ones selected in D1. The topological modularity of the core models generated with redGEM eased the transferability of steady-states across models, allowing us to preserve similar values for fluxes and concentrations for the overlapping reactions of the three models.

We connected the core models to the biomass building blocks (BBBs), as defined by Neidhart *et al.* [78], via lumped reactions generated with the lumpGEM algorithm by applying approaches developed by Ataman and Hatzimanikatis [58]. A lumped reaction is a reaction that collapses a subnetwork of reactions into one mass-balanced reaction. D1-3 had 247, 189 and 196 lumped reactions, respectively. The models' lumped reactions are indeed not the same across D1-3. Consequently, lumped reactions impose certain stoichiometric constraints that can require flux to pass through alternative metabolic routes within the models. For instance, a BBB can be produced by a completely different lumped reaction (Figure 2.3A), as we can generate it via a different subnetwork of reactions in the systems with larger cores. Thus, having distinct lumped reactions results in the redistribution of the flux profiles across models. An example of this is the hub metabolite methylglyoxal that provides new alternatives for lumped reactions in D2 and D3, thus contributing to differences in flux distribution across the models.

We studied the lumped reactions and in D1-3 and observed that 103 were common between the models (Figure 2.3B). D1, D2 and D3 have 126, 57 and 66 lumped reactions that are unique to themselves. D1 requires considerably more lumped reactions in order to produce the BBBs from the core subsystems. If we consider the lumped reactions as subnetworks of reactions, 474, 453 and 458 reactions are used to build the lumped reactions of D1-3, respectively (Figure 2.3C). Interestingly, 446 reactions are common between the pools of reactions that constitute the lumped reactions of D1-3.

In order to ensure equivalency between D1-3, we proposed a procedure that uses a Mixed Integer Linear Programming (MILP) formulation that imposes similarity between the representative steady states of the models (Methods). The D2 fluxes of central carbon reactions are within below one percent deviation from the reference flux of D1 (Appendix A3). The only central carbon fluxes in D3 that deviate from D2 reference flux with more than one percentage are transaldolase (TALA) and xylose isomerase (XYLI2) with 4.5% and 33.2% respectively. The concentration profile of D2 is within one percent of D1 reference steady state, except for ADP, CoA, S-dihydroorotate and L-glutamine with 16%, 45%, 303% and 94% deviations from D1. On the other hand, the D3 metabolite concentration steady-state is within one percentage from the D2 metabolite concentration vector. The modularity and the consistency of redGEM and lumpGEM algorithms in GEM reduction allowed the steady-states to be transferred and communicated between models efficiently.

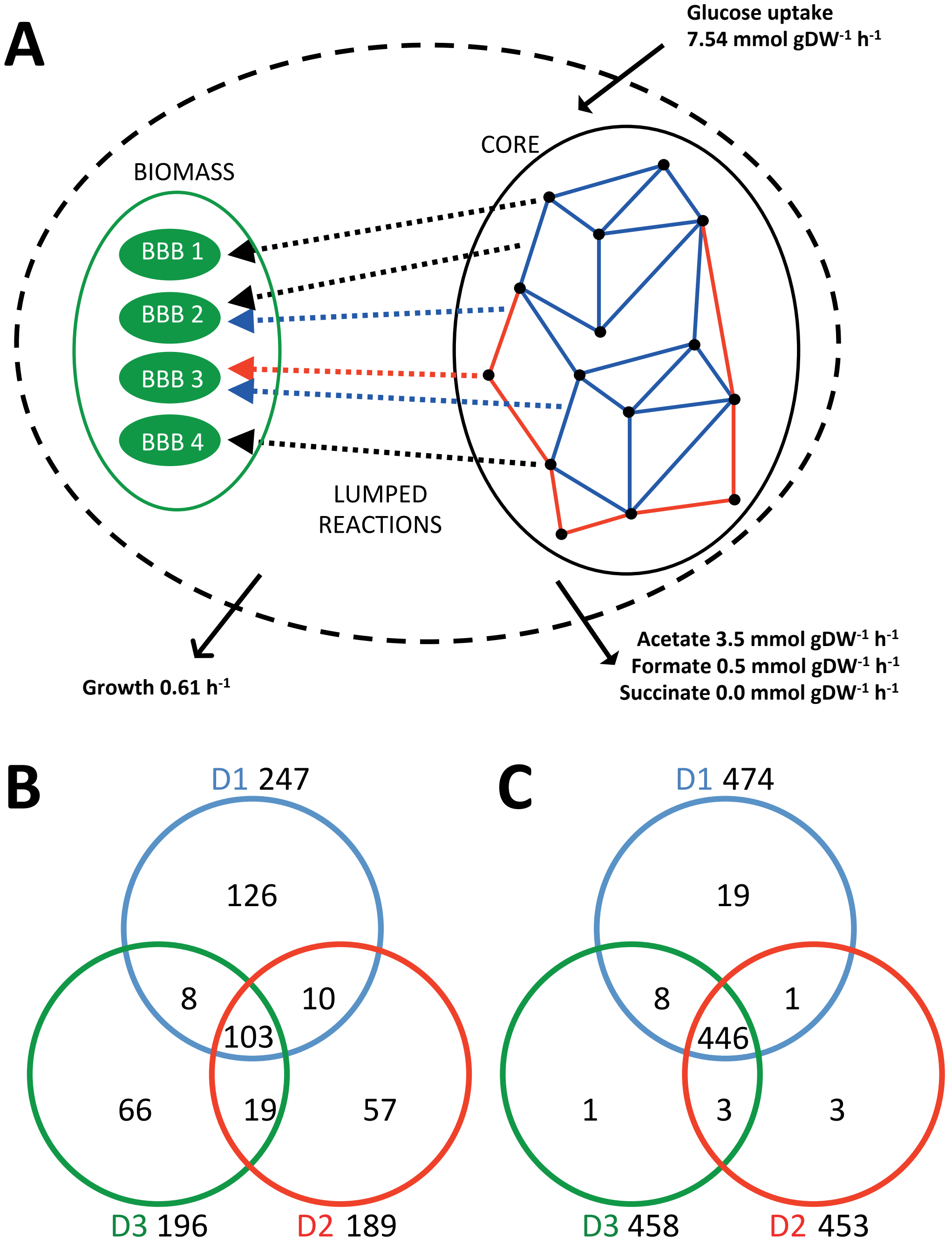


Figure 2. 3. Illustration and analysis of core metabolism connections to biomass building blocks via lumped reactions. The schematic representation (A) of the studied metabolic networks where the edges are reactions and the nodes are metabolites. The blue edges and nodes belong to the core network of D1. The red edges and nodes correspond to additions from the D2 expansion. The dashed arrows are lumped reactions of the system, connecting the core to the biomass, where: black represents lumped reactions that are present in all the models, blue is for lumped reactions existing only prior to the expansion in D1 and red is for lumped reactions existing only after network expansion in D2. Biomass building blocks (BBBs) are represented by the green ovals. The black solid arrows indicate fluxomics data that were integrated for optimally grown *E.coli* [61]. Venn diagrams

highlight differences in the lumped reactions of D1, D2 and D3 in terms of (B) subnetworks and in terms of (C) reactions.

2.2.3.2. Equivalence in kinetic parameters

We constructed kinetic models around the selected reference steady-states of D1-3 using the ORACLE workflow [50, 84-86]. Uniform Monte Carlo sampling of the degrees of saturation of the enzyme active sites allowed us to study the kinetic parameter space, as proposed by Wang and colleagues [84]. The local stability of the models generated was tested by verifying that the eigenvalues are not positive. We first sampled 50’000 stable kinetic models for D1. To ensure equivalency at kinetic parameter level between D1-3, we adapted the ORACLE workflow to allow fixing the sampled saturation states from one model to another (Methods). From the 50’000 stable D1 kinetic models, we found 96.1% (48’080) to be stable in D2, of which 98.4% (47’299) were stable in D3. We then computed the MSCs for these stable models in order to compare how MCA-based decisions are affected by metabolic network size.

2.2.4. Consistency in MCA across models

2.2.4.1. Ranking enzymes for flux control

Some fundamental cellular tasks for a given physiology include metabolite excretion, substrate uptake and cellular growth, μ. As we studied the physiology of optimally grown *E.coli*, we considered control over μ across models to assess the consistency in conclusions based on MCA outputs. The flux control coefficients (FCCs) of μ were ranked for D1-3 based on their absolute means across stable models. The models were compared pairwise in increasing order of size (i.e. D1 versus D2, and D2 versus D3) to assess the impact of systematic network expansion on MCA (Figure 2.4).

The cellular growth FCCs with respect to PGI, phosphofructokinase (PFK) and ATP maintenance (ATPM) are the most consistent in terms of sign and magnitude when comparing D1 with D2 (Figure 2.4A). Pyruvate kinase (PYK), fructose biphosphate aldolase (FBA) and 2-oxogluterate dehydrogenase (AKGDH) are also in agreement in terms of sign but magnitude can differ significantly. Some enzymes have control in D1 but no control in D2, such as ribulose 5-phosphate 3-epimerase (RPE) and phosphoglycerate mutase (PGM). Others, vice versa, have control in D2 but no control in D1 such as phosphoglycerate kinase

(PGK) and glucose 6-phosphate dehydrogenase (G6PDH2r). Ribose-5-phosphate isomerase (RPI), on the other hand, has opposing control on cellular growth in the two models. Differences in FCCs of cellular growth between D1 and D2 suggest that the modular expansion of D1 to D2 significantly affects the control scheme.

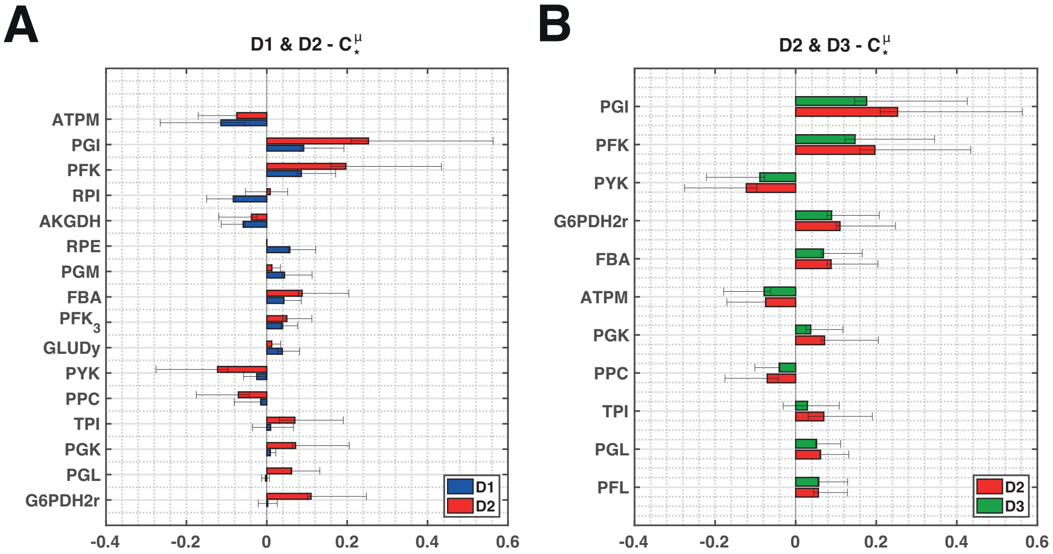


Figure 2. 4. Cellular growth (μ) control across models. Pairwise illustration of the union of the top 7 enzymes across the models in terms of absolute control over cellular growth for (A) D1 versus D2, and (B) D2 versus D3. The whiskers give the upper and lower quartiles of the FCC populations and the bars give the means.

We then compared top cellular growth FCCs in D2 and D3, which are in great sign and magnitude agreement (Figure 2.4B). PGI, PFK and PYK are the top three enzymes in terms of cellular growth control according to both D2 and D3. The consistency between these FCCs suggests that the modular expansion of D2 to D3 does not affect the control pattern as significantly as the network expansion from D1 to D2. An analogous analysis was carried out for the flux control of glucose uptake and several other cellular excretions (Appendix A4), and we observed a similar trend. The differences in control patterns appear to be significant when expanding from D1 to D2, but of lesser importance when expanding from D2 to D3. This finding could suggest that entire genome-scale kinetic models are not necessary to capture the essential physiological features of a cell as long as the model reduction is done systematically around carefully selected subsystems of study interest. Additionally, this could mean that D1 is possibly missing on some important information for performing MCA.

Clearly, a study-specific resolution criterion in terms of model size/complexity that has to be met needs to be established before a model is used for further analysis.

2.2.4.2. MCA consistency across reduced models

As the study above revealed, certain flux control patterns can change significantly between models due to network complexity. We tried to locate, analyze and understand the differences and the similarities in MSCs that occur due to the topological alterations in kinetic model complexity. According to MCA theory, the FCCs conform with the summation theory [87, 88]. We proposed a deviation index (DI) that provides a quantitative measure on how much a reaction's FCCs differ between two models, postulated from the summation theory (Methods). The DI served as a metric to classify reactions with respect to their consistency in FCCs across the reduced models.

We estimated the DI of 271 common enzymatic reactions when expanding from D1 to D2 to predict deviations in FCCs for the system. Reactions with the lowest DI (0-25 percentile) were mostly from the central carbon metabolism (Figure 2.5). The reactions with the highest DI (75-100 percentile) were mostly located in the ETC. The only central carbon metabolism reactions having a high DI were TALA, acetyl-CoA synthase (ACS), phosphoenolpyruvate synthase (PPS) and NAD malic enzyme (ME1). TALA produces D-fructose 6-phosphate and, PPS and ME1 involve transformation of pyruvate. D-fructose 6-phosphate and pyruvate are both central carbon metabolites around which the expansion adds reactions (Figure 2.2 and 2.5). ACS is only one reaction away topologically from pyruvate, around which the expansion adds a reaction (Figure 2.2 and 2.5).

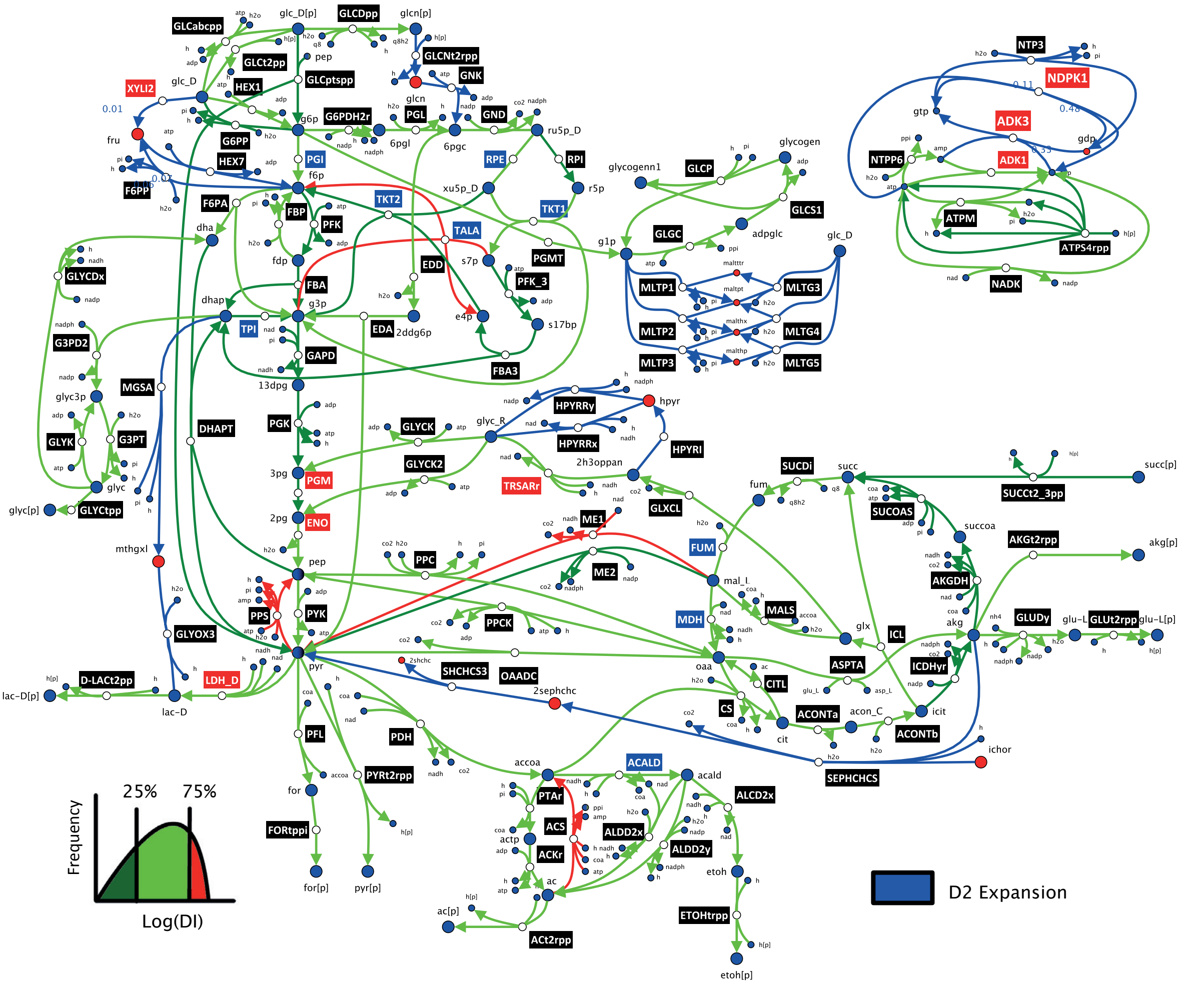


Figure 2. 5. *E.coli* network diagram illustrating the deviation index (DI) of reactions when scaling up from D1 to D2. The reactions added by the modular redGEM expansion from D1 to D2 are given in blue and the 271 enzymatic reactions in common between D1 and D2 are color-coded based on the DI. The reactions were categorized into low DI (0-25 percentile), medium DI (25-75 percentile) and high DI (75-100 percentile), and are given in dark green, light green and red, respectively. Diagram does not include all the reactions of the systems.

We repeated the above analysis for D2 and D3, where we analogously compute the DIs for the 307 common enzymatic reactions (Methods). Similar observations were made for the reactions having low DIs (0-25 percentile) as most were located in central carbon metabolism, within the subsystems around which reduction was performed (Figure 2.6). The reactions with higher DIs (75-100 percentile) are predominantly located around the ETC, with the exception of several reactions pertaining to central carbon metabolism. As with the previous analysis of D1 versus D2, ME1 and ACS had high DIs. The D3 expansion adds reactions around pyruvate (Figure 2.2 and 2.6), which could explain this observation. Aspartate transaminase (ASPTA), phosphoenolpyruvate carboxykinase (PPCK) and succinate dehydrogenase (SUCDi) from the central carbon metabolism exhibited high DIs (Figure 2.6). ASPTA is directly connected via 2-oxoglutarate with NADPH glutamate synthase (GLUSy), which is a newly added reaction by the D3 expansion. PPCK is connected via polyenolpyruvate to 3-phosphoshikimate 1-carboxyvinyltransferase (PSCVT), another add by the D3 expansion. Furthermore, SUCDi is topologically connected with the added reaction ubiquinone L-Lactate dehydrogenase (L-LACD2), as cofactors ubiquinone-8 and ubiquinol-8 partake in both reactions. Interestingly, periplasmic glucose dehydrogen (GLCDpp), where ubiquinone-8 and ubiquinol-8 also participate, has a high DI as well. GLCDpp possibly causes its neighbouring reactions gluconokinase (GNK) and D-gluconate transport (GLCNt2rpp) to have high DIs too, due to stoichiometric coupling. These observations suggest that alterations in flux split ratios around important branching points - caused by network expansion - could result into higher DIs in reactions at their vicinities.

Overall, lower DIs were observed for reactions having a higher flux, pertaining to the core central carbon metabolism around which the models D1-3 were reduced (Figures 2.5 and 2.6). Since the cores of the reduced models contain the 12 precursor metabolites for biomass, their control patterns were expected to be similar. Stephanopoulos and Vallino point out that metabolic pathways of organisms have evolved over time to resist flux alterations at branching points [89]. The control architecture of an organism is built such that it preserves the flux splitting ratios of essential metabolic nodes. However, if two models have differences in the number of reactions and/or in the flux splitting ratios around an important branching point, the control architecture of the two systems can differ considerably.

Since we studied optimally grown *E.coli*, it was expected that the D1 to D2 expansion with the addition of XYLI2, F6PP and HEX7 would have influence on control patterns: the flux splitting ratios around the essential biomass precursor D-fructose 6-phosphate is altered. D-fructose 6-phosphate is a critical metabolic node for producing cell wall biomass building blocks and is located relatively upstream in the process of glucose catabolism. Altering flux splitting ratios around D-fructose 6-phosphate will have direct implications on the fate of the carbon flow across the whole network, particularly due to its upstream location.

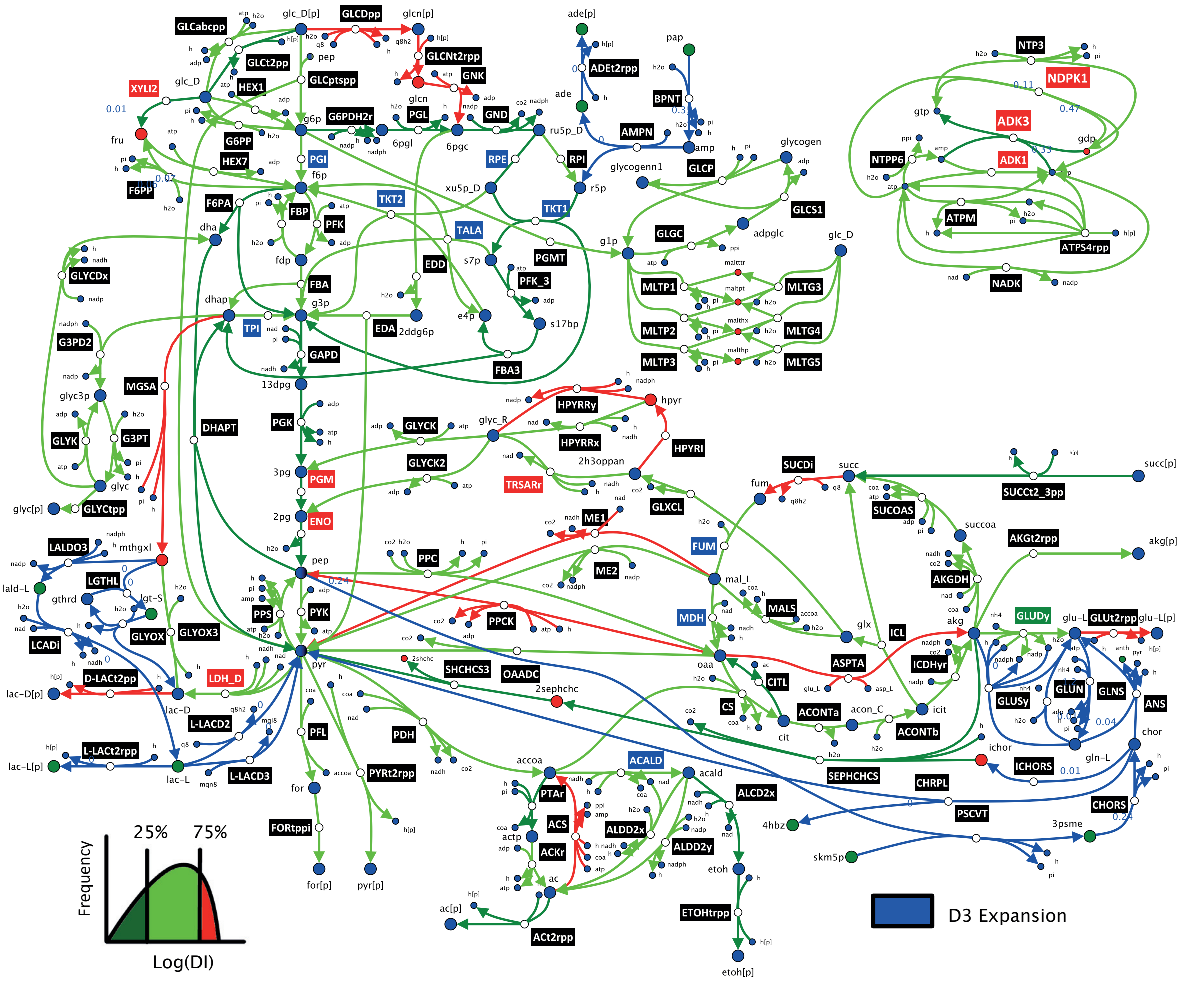


Figure 2. 6. *E.coli* network diagram illustrating the deviation index (DI) of reactions when scaling up from D2 to D3. The reactions added by the modular redGEM expansion from D2 to D3 are given in blue and the 307 enzymatic reactions in common between D2 and D3 are color-coded based on the DI. The reactions were categorized into low DI (0-25 percentile), medium DI (25-75 percentile) and high DI (75-100 percentile), and are given in dark green, light green and red, respectively. Diagram does not include all the reactions of the systems.

The expansion from D2 to D3 results in different flux splitting ratios around three biomass precursors: pyruvate, polyenolpyruvate and 2-oxoglutarate. Again, we can expect flux control patterns across the models to differ as the proportion of carbon flow directed towards certain biomass building blocks is affected. However, within the central carbon metabolism, these precursors are located relatively downstream to the glucose uptake, compared to for instance D-fructose 6-phosphate. Consequently, we can expect that these flux splitting ratios have less impact on the growth control of the system than D-fructose 6-phosphate. If we were discussing the production of certain amino acids of interest, rather than just cellular growth, these ratios could be of higher importance to the analysis. The significance of a metabolic node is strongly subject to the scope of the study. Hence, it is difficult to imagine a “one-size-fits-all” model due to the complexity of the problems encountered in metabolic engineering.

Indeed, the importance of a metabolic branching point is very study-specific as objectives can vary significantly. Had we, for instance, been interested in the study of D-lactate production, it would have been essential to include the metabolism of methylglyoxal, D-lactate and L-lactate into the subsystems around which model reduction is performed. However, as we are not interested in the production of D-lactate, we are not that concerned about the high DI of D-lactate transporter (D-LACt2pp) when comparing D2 and D3 (Figure 2.6). Furthermore, if we were interested to produce D-lactate, it would be essential to consider implication of attempting to deviate flux towards the metabolism of D-lactate. If the redirection of flux towards D-lactate imposes important changes in the flux splitting ratios of significant metabolic nodes of wild-type *E.coli*, it may be worth considering other organisms that cause fewer modifications in flux distribution [89, 90].

2.2.4.3. Study of uncertainty in MCA

The MCA results of D1-3 were further studied by comparing their absolute deviations in the FCCs. We considered with respect to the central carbon subsystems to find which central carbon enzymes contributed in most uncertainty across the networks. The FCCs of reactions in the glycolysis (Figure 2.7) appear to have most absolute deviation stemming from enzymes in the glycolysis and in the PPP. When comparing D1 with D2 (Figure 2.7A), enzymes in glycolysis contribute most to the uncertainty whereas in the comparison of D2 and D3 (Figure 2.7B), the PPP contributes the most. Again, the additional connections around D-fructose 6-phosphate (Figures 2.1 and 2.2) when expanding from D1 to D2 could explain this. Differences in flux splitting ratio around D-fructose 6-phosphate affect the redistribution of the flux in the network and hence the control pattern. Generally, reactions with a larger flux exhibit less absolute deviations in their FCCs. This parallels the observation that central carbon reactions carrying higher flux are perhaps more rigid in control patterns.

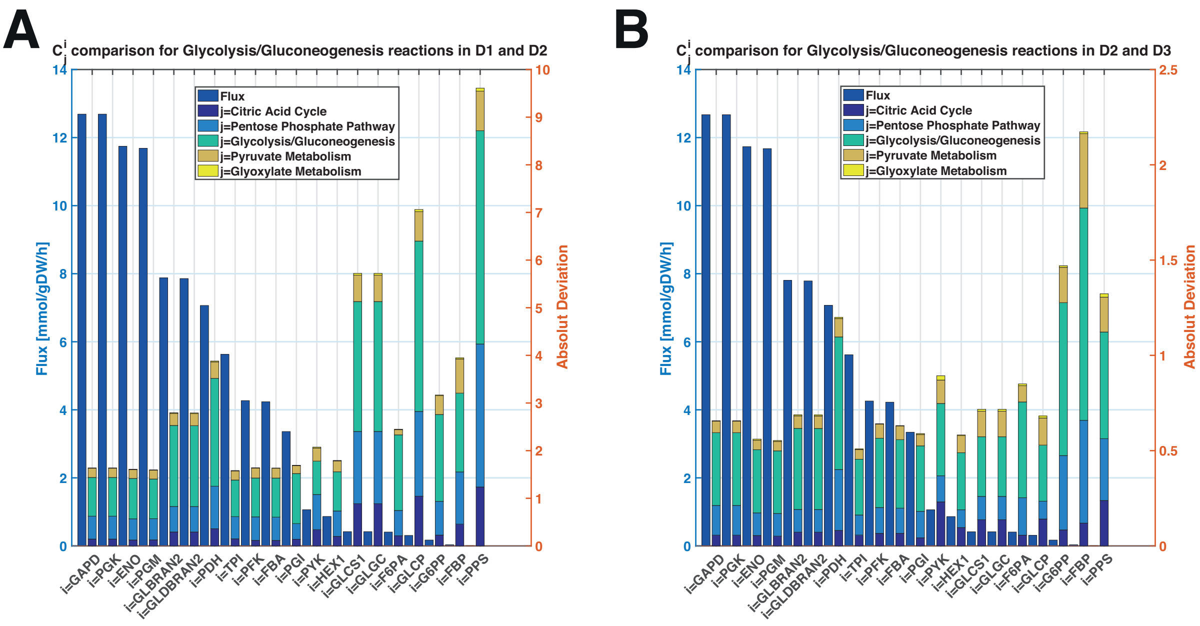


Figure 2. 7. Comparison of flux and absolute deviations in FCCs for glycolytic reactions for (A) D1 versus D2, and (B) D2 versus D3. The absolute deviations computed per subsystem correspond to the sum of the absolute deviations in FCCs of reaction i with respect to all enzymes of the subsystem j. The reactions contained in a subsystem are as defined in the original GEM that was reduced [77]. The flux values shown did not deviate by more than 1% between pairs of models.

We perform a parallel analysis on FCCs of PPP reactions and similar observations were made. The expansion from D1 to D2 has most absolute deviations coming from enzymes from glycolysis (Figure 2.8A), whereas, the expansion from D2 to D3 has most deviations due to PPP enzymes (Figure 2.8B). Again, reactions carrying higher flux have less absolute deviation in their FCCs between the pairs of models. We analyzed FCCs individually in terms of absolute deviation (Appendix A5), for both pairs D1 and D2 as well as, D2 and D3. PGI, TPI and PFK were the top three central carbon enzymes that resulted in the most absolute difference in flux control across the network. From the PPP enzymes, RPI resulted in the most absolute deviation in flux control. We also recall that RPI had sign-wise opposing control on cellular growth in the comparison of D1 and D2 (Figure 2.4A). Due to the highly non-linear nature of the studied systems, it is difficult to make direct conclusions on the causality of the observed deviations in control patterns of the networks. Most of the deviations were observed amongst peripheral transport reactions, rather than central carbon metabolism (Appendix A5). Nevertheless, we could still find metabolic engineering decisions relevant to our study, independent of the complexity based on MCA outputs (Figure 2.4).

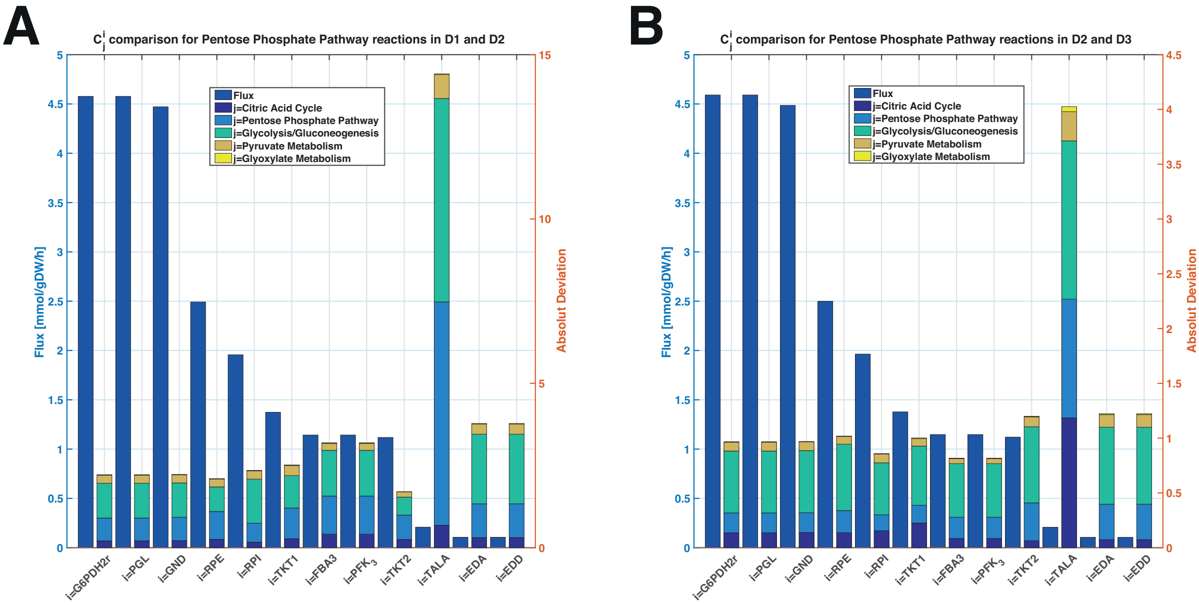


Figure 2. 8. Comparison of flux and absolute deviations in FCCs for PPP reactions for (A) D1 versus D2, and (B) D2 versus D3. The absolute deviations computed per subsystem correspond to the sum of the absolute deviations in FCCs of reaction i with respect to all enzymes of subsystem j. The reactions contained in a subsystem are as defined in the original GEM that was reduced [77]. The flux values shown did not deviate by more than 1% between pairs of models.

2.3. Conclusion

This work studied the impact of model complexity on the metabolic engineering decisions derived from MCA outputs. The redGEM and the lumpGEM algorithms were used to consistently reduce the *E.coli* iJO1366 GEM. Omics data for the physiology of optimally grown *E.coli* was integrated into the reduced stoichiometric models. The modularity of the reduced models assisted us in the development of a workflow allowing to preserve maximum equivalence between the flux profile and metabolite concentration steady states. The ORACLE framework was used to generate populations of stable kinetic models around these reduced stoichiometric models. Our workflow ensured that we preserve equivalency amongst the populations of the kinetic parameters for the stable kinetic models. The MSCs were computed within the ORACLE framework for the populations of stable kinetic models. Analysis of the MSCs, revealed that we can derive context-specific metabolic engineering conclusions that are independent of the model's complexity, as long as the reduction is performed consistently.

The “usefulness” of a kinetic model is highly dependent on the objectives of the study being undertaken. We selected the subsystems for the GEM reduction such that we: (1) cover the essential biomass precursor metabolites according to Neidhart as we focused primarily on cellular growth control and, (2) that we capture the ETC essential to account for redox potentials. The addition of reactions around D-fructose 6-phosphate when expanding from D1 to D2 appeared to significantly affect growth control patterns (Figure 2.4A). However, the expansion from D2 to D3 had considerably less impact as top cellular growth FCCs are consistent (Figure 2.4B). As D-fructose 6-phosphate is an essential precursor for cell wall fabrication, a network expansion affecting flux distribution around it can be expected to have significant impact on cellular growth control structure. Hence, it is essential to consider the importance of certain metabolic nodes with respect to the study goals in order to ensure no information is lost in the reduction. Again, importance of a metabolic node is strongly influenced by the nature and the objectives of the analysis.

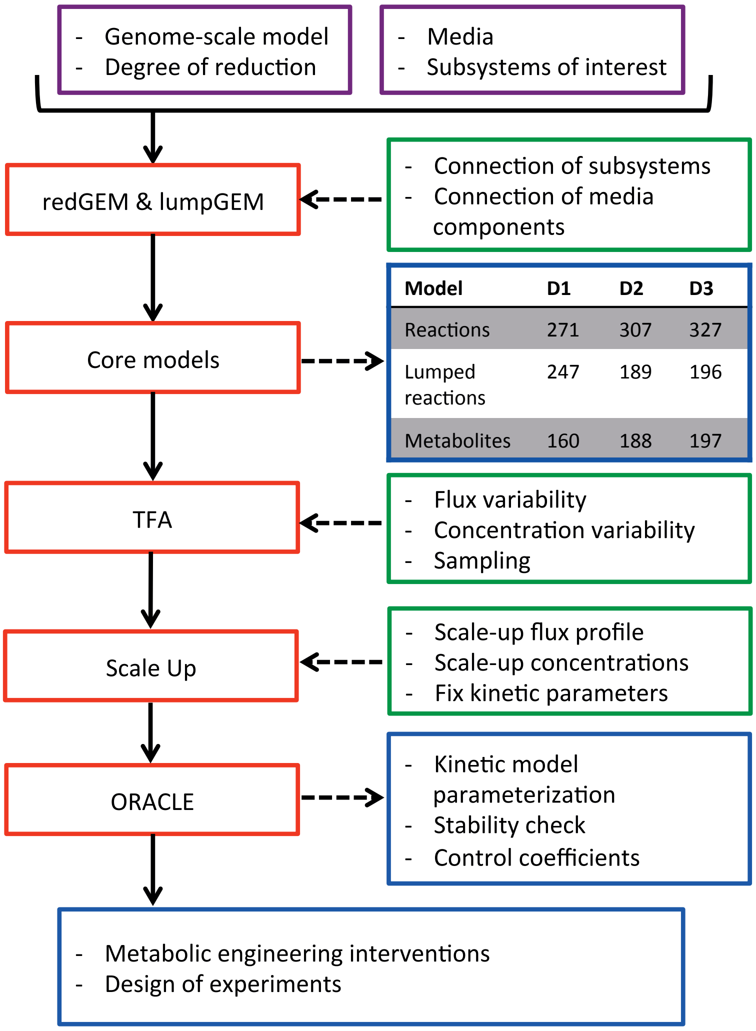
The MCA summation theorem was used to postulate a deviation index (DI) that gave a numerical indication on the consistency of the FCCs with respect to a reaction. Most of the reactions around central carbon metabolism, carrying a higher carbon flux, appeared to have lower DIs. Flux control for reactions with larger fluxes were more robust, particularly if the number of connecting reactions did not change between models for the metabolites participating in the reaction. The larger DIs were noted in the ETC and peripheral reactions. Nevertheless, the consistency in the control patterns across reduced models allows us to make conclusions that are independent of the network complexity. In fact, we may not need full genome-scale kinetic models when the model reduction is done consistently as the essential, study-specific, information is accounted for by the reduced model.

This chapter demonstrates that systematic and modular model reduction algorithms ease the scale-up of kinetic models of metabolism. Our workflow describes an MILP formulation for insuring maximum equivalence between models when transferring steady-states. Furthermore, the workflow ensures that the kinetic parameters are kept as similar as possible between the populations of stable kinetic models built around the reduced stoichiometric models. To our knowledge, this is the first effort to date demonstrating transferability of steady-states between large-scale kinetic models, whilst obtaining consistent, study-specific, metabolic engineering decisions. As systematic model reduction algorithms gain momentum in the field, we hope to pave a path towards building more robust and transferable kinetic models for the community.

In this chapter, we studied how steady-states can be transferred from one kinetic model to another one of different complexity level. However, it is known that due to the underdetermined nature of the biological system, multiple steady-state solutions could characterize the physiology. In the next chapter, we will discuss the importance of considering alternative steady-states when constructing kinetic models, as they can significantly affect conclusions drawn from model outputs.

2.4. Materials and methods

We developed a workflow for building consistently reduced kinetic models from a genome-scale metabolic model (Figure 2.9). We used the redGEM algorithm to construct core models of increasing network size from the *E.coli* iJO1366 genome-scale model. The lumpGEM algorithm was used to generate lumped reactions for the biosynthesis of biomass building blocks (BBBs) for these models. We used thermodynamic-based variability analysis (TVA) [83] to study the flexibility of the models. We proposed a procedure for scaling up the flux and concentration steady-states from one model to another one using the MILP formulation. The ORACLE framework was enhanced, allowing us to keep parametric equivalency between the populations of kinetic models around the steady states of the reduced models. These steps are further detailed below.



Model	D1	D2	D3
Reactions	271	307	327
Lumped reactions	247	189	196
Metabolites	160	188	197

Figure 2. 9. Workflow for building consistent reduced kinetic models. The various steps are discussed with further detail within the chapter.

2.4.1. Model reduction

The stoichiometry of the core networks was defined with the redGEM algorithm, which reduces systematically genome-scale model reconstructions of metabolism [57]. The *E. coli* iJO1366 genome-scale model was reduced, with aerobic minimal media, glucose as the sole carbon source, and the selected starting subsystems corresponding to central carbon metabolism (glycolysis/gluconeogenesis, citric acid cycle, pentose phosphate pathway, pyruvate metabolism, and glyoxylate metabolism). We incorporated all the reactions that use metabolites of the quinone/quinol pools (ubiquinone, ubiquinol, menaquinone, menaquinol, 2-dimethyl menaquinone and 2-dimethyl menaquinol) as the electron transport chain subsystem in order to account for the energy metabolism of the system. redGEM allows the user to define a degree of connection, D, to define the level of connectivity of the core. D is an input parameter of the redGEM and lumpGEM. D corresponds to the number of reaction required to connect the pairs of metabolites between starting subsystems, as defined in redGEM [57]. We generated core networks with a D of 1,2 and 3, which gave rise to models D1, D2 and D3 respectively. The lumpGEM algorithm [58] was used to generate lumped reactions for the biosynthesis of the BBBs for these core networks. Lumped reactions are sub-networks of reactions composed of non-core reactions that can be used to produce a BBB. Alternative lumped reactions of the minimal sub-network size were kept for each of the BBBs. Reactions that could not carry flux were considered as blocked and were removed.

For some of intracellular metabolites, a corresponding transport reaction has not been biochemically characterized and does not appear in the *E. coli* iJO1366 and in our reduced model. However, these metabolites, unless they are highly polar or very large, are subject to passive diffusive transport through the cell membrane. Therefore, we explicitly added transport reactions for these metabolites that operate at least at basal level (10^-6 mmol/(gDW*h)).

2.4.2. Flux directionality assumptions

We make the following directionality assumptions for several bi-directional reactions:

- Fructose-biphosphate aldolase (FBA) that is part of mid-lower glycolysis is set towards catabolism [79].
- The bi-directional transports of magnesium and phosphate are both set towards uptake [80, 81].
- Acetate kinase (ACKr) and phospho-transacetylase (PTAr) are both set towards the acetate production, because acetate is one of the main by-products [61].
- The succinyl-CoA synthetase (SUCOAS) is set towards the production of succinate [61].

The polyphosphate kinases (PPK2r, and PPKr) are set towards the polyphosphate polymerization [82].

2.4.3. Thermodynamic analysis

The available fluxomics and metabolomics data for the optimal growth of *E.coli* under aerobic conditions and minimal media was integrated in our models. The MILP formulation of the thermodynamics-based flux analysis was used to implement these data into D1, D2 and D3. Since the models were used to build kinetic models, it was undesirable for reactions to be at thermodynamic equilibrium, which would result in them having equal backward and forward fluxes. We imposed MILP constraints to ensure that the thermodynamic displacement, Γ [52, 84, 87], is not at equilibrium. For reactions near equilibrium $\Gamma \approx 1$.

2.4.4. Maximum equivalency between steady-states

We sampled the flux space of D1 in order to characterize the solution space without violating physiological, thermodynamic and directionality constraints. The convexity of the solution space enabled us to efficiently sample using the Artificial-Centering Hit-and-Run sampler in the COBRA Toolbox [8, 9]. We sampled flux vectors and used Principal Component Analysis (PCA) [91] to select a mean reference state. Similarly, we sampled and selected the reference state with PCA for the concentration solution space of this selected flux profile.

In order to make the comparison of the models equitable, we wanted to maintain most similar steady states between the models. For instance, for D2 we would like the flux vector to be the equal possible to the one from D1. Topological differences in the models make it impossible to have numerically exactly the same flux distribution in larger model for the same reactions. Hence, we take the representative flux from D1 and apply it with percentage relaxation with upper and lower bounds, $F^{ub}_{rxn,i}$ and $F^{lb}_{rxn,i}$ respectively, into D2. Consequently, we use an MILP formulation to minimize the number of violations of flux boundaries that we are trying to impose. For each intracellular reaction that is shared between the two models we create a binary variable $z_{rxn,\,i}$ so that when it is equal to 1, the constraints that we impose become inactive. We add for each of these reactions the following constraints:

$$NF_{rxn,i} + (F^{ub}_{rxn,i} - UB_{rxn,i}) * z_{rxn,\,i} < F^{ub}_{rxn,i}$$

$$NF_{rxn,i} + (F^{lb}_{rxn,i} - LB_{rxn,i}) * z_{rxn,i} > F^{lb}_{rxn,i}$$

where, *UB* and *LB* are the upper and lower bounds of the net fluxes *NF* of the reactions. We minimize the sum of the binaries, $z_{rxn,i}$, in order to have minimal violation of the flux constraints:

Minimize:

$$\sum_{i}^{\#Fluxes} z_{rxn,i}$$

Subject to:

$$S.v = 0$$

We implied a 1% relaxation to apply and test how many flux constraints we can impose without violation (minimal number of active binary variables $z_{rxn,i}$). After applying the constraints that are not violating model boundaries of D2, we proceed to sampling the solution space. We selected a sample based on mean PCA as with the representative flux of D1. We then implied in a similar manner the concentration profile from D1 into D2 with a 1% relaxation and sampled the concentration space for this flux profile. We repeat this procedure when scaling up the flux and concentration steady states from D2 into D3.

2.4.5. Constructing kinetic models

We used the ORACLE framework to build 50'000 kinetic models around the steady states for D1, D2 and D3. Available kinetic properties of enzymes from the literature [92] and the databases [93, 94] were incorporated. Reversible Hill kinetics [95] and convenience kinetics [96] were used for reactions with unknown kinetic mechanism (Appendix A6). Kinetic mechanisms with no or partial information about their parameter values were sampled within the space of kinetic parameters in the form of degree of saturation of enzyme [84]. We parameterized a population of kinetic models and performed consistency tests [51, 54, 84]. We then computed the flux and concentration control coefficients [84, 97]. For further details on the ORACLE workflow the reader is referred to [48, 50-52, 84-86, 98].

We preserved equivalency between populations of kinetic models for D1-3 by fixing the degree of saturation of enzymes from less complex models into the more complex models. We wanted to preserve model equality so that we can fairly compare MCA outputs of the models. Within the ORACLE framework, we added a feature for fixing the degree of saturation of enzymes. For the parameters that were common between D1 and D2, we fixed the degrees of enzyme saturations from D1 models into D2 models and we sampled the rest of the D2-specific parameters uniformly, until we found a stable model. A maximum of 1’000 trials were made to obtain a stable model. Hence, we preserved equivalency of the kinetic parameters between D1 and D2. Analogously, we repeated this procedure to imply the degrees of enzymes saturations from D2 into D3.

2.4.6. Control coefficient deviation index

In MCA the FCCs conform with the summation theorem defined in [87, 88]. The theorem implies that all the metabolic fluxes are systemic properties of the model and that their control is shared by all the reactions within the system. The summation theorem makes the assumptions that: (1) the parameters for which we compute flux control coefficients are of first order with respect to the flux, and that (2) the sum of a flux’s control coefficients with respect to all the parameters of the system is equal to one. We proposed a deviation index (DI) derived from the summation theorem to quantify the discrepancies in control patterns of a flux between two different models (Figure 2.10).

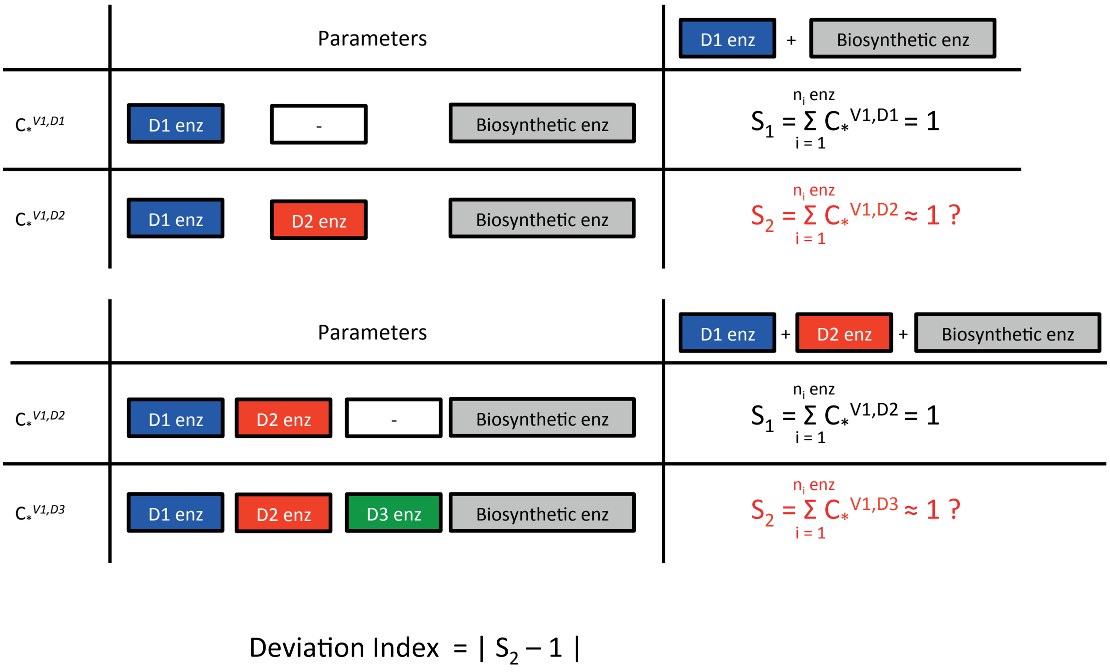


Figure 2. 10. Derivation of the deviation index (DI) from the summation theorem.

2.5. Appendix A

A1 Table. Fluxomics and metabolomics data incorporated in the model. Table with the fluxomics data in mmol/gDW/h and the concentration data in log(M).

A2 Table. Thermodynamics-based variability analysis of models. Spreadsheet with the list of metabolites and reactions inside the models with variability analysis of metabolic flux and metabolite concentrations.

A3 Table. Flux and concentration steady states. Spreadsheet providing metabolic flux and concentration reference steady-states across the models with a comparative study.

A4 Figure. Flux control coefficients of (a) glucose uptake, (b) formate excretion and (c) acetate excretion across the models. Pairwise illustration of the union of the top 7 enzymes across the models in terms of absolute control over cellular growth for D1 versus D2, and D2 versus D3. The whiskers give the upper and lower quartiles of the FCC populations and the bars give the means.

A5 Table. Analysis of absolute deviations in means of flux control coefficients for the entire systems. Further assessment of deviations in flux control coefficients between model expansions from D1 to D2 and from D2 to D3.

A6 Supporting information. Kinetic mechanisms used for the models.

3. Kinetic models of metabolism that consider alternative steady-state solutions of intracellular fluxes and concentrations

3.1. Introduction

Over the last decades, advances in genome editing technologies have allowed the redirection of carbon flow within the organism towards specialty products of interest and desired physiologies [32]. Identifying candidate enzymes is fundamental for genetic modifications that have seen applications in metabolic engineering, basic and applied biology, biotechnology and medical sciences [99-101]. Increasingly available high-throughput sequencing data has enabled the construction of stoichiometric genome-scale metabolic models (GEMs) that describe mathematically the balanced metabolic fluxes within an organism [2]. Metabolic models such as these GEMs have been extensively used to characterize overall network behavior of organisms, which can provide guidance about the genes that can be modified to improve a desired product biosynthesis. Improved guidance for metabolic engineering and basic biology will be achieved with kinetic models of the reactions/networks in GEMs.

The construction of a kinetic model of metabolism requires knowledge of steady states and/or dynamics of metabolic fluxes and metabolite concentrations that can be used to estimate the unknown kinetic parameters that describe these data. However, there are many sources of uncertainty in metabolic fluxes and metabolite concentrations that result in partial knowledge. Advances in C13 isotopomer techniques facilitated the measurement of fluxes across cellular reactions and promoted the development of metabolic flux analysis (MFA) [102]. One main uncertainty in fluxes is the flux directionality as reactions can be thermodynamically bidirectional [16]. Metabolomics and thermodynamics can be used as it is done in thermodynamic-based flux analysis (TFA) [14-16] to constrain the direction of some of these fluxes. But even when information about the directionality of all the reactions and fluxomics from labeling experiments are used, there is still a great uncertainty on exact estimation of fluxes as the degrees of freedom remain high, especially as we increase the size of the networks. The addition of constraints based on measured gene expression data [19, 103] and enzymatic data [25] can reduce the degrees of freedom. However, the system remains underdetermined, resulting in multiple alternative steady-state flux distributions corresponding to the physiology under study. Different steady-state solutions could directly affect the predictions of kinetic models, leading towards very distinct conclusions and guidance for metabolic engineering.

Several promising methods exist for constructing kinetic models around representative steady states of metabolic fluxes and metabolite concentrations [44, 54]. The Optimization and Risk Analysis of Complex Living Entities (ORACLE) workflow [50, 51, 84] and frameworks built around ensemble modeling [45, 47] have made significant strides towards genome-scale kinetic modeling of metabolism. These methods generate populations of non-linear kinetic models around a selected reference steady state (RSS) that is chosen based on its ability to characterize the observed physiology. Methods commonly used for selecting a RSS include using the computed optimal solution to an objective function that defines physiological tasks [104], fitting the data from MFA [102], or performing principal component analysis (PCA) on a sampled solution space [50]. Once a RSS is established, kinetic models are constructed around it, which allows the study and prediction of cellular metabolic response to perturbations [105]. These populations of kinetic models can be studied using statistical procedures to identify target enzymes, sensitively analyze kinetic parameters, and design experiments [48, 49]. There is no unique and evident approach for selecting a RSS for such an underdetermined system. To our knowledge, the impact of alternative RSSs describing a physiology using the kinetic parameters and the outputs of these kinetic models have not been studied.

Hereby, this chapter examines how uncertainty in intracellular flux solutions and metabolite concentrations influences the metabolic control analysis (MCA) of populations of non-linear kinetic models built around alternative steady states. We integrated physiological data from *E. coli* grown aerobically in a batch cultivation [61] into a reduced core model derived from the iJO1366 *E. coli* GEM [57, 58, 77] and found that the data were not sufficient to uniquely determine the steady state metabolic flux distribution as several reactions could operate in either the forward or reverse direction. These so-called bi-directional reactions result in the existence of multiple feasible flux directionality profiles (FDPs) that represent the same physiology, because in any FDP, reactions operate only in one direction [14]. We constructed populations of kinetic models for 4 selected FDPs to demonstrate how significantly MCA outputs and metabolic engineering decisions are affected.

3.2. Results and discussion

The procedure for characterization and analysis of steady-state multiplicities arising from the underdetermined nature of the system is a constitutive part of the ORACLE workflow [48, 50, 51, 84-86, 98]. The workflow assists with more reliable and robust MCA-based metabolic engineering decisions that will enable the identification of study-specific target enzymes, independent of the steady state. Various types of biological data are combined into a thermodynamically feasible stoichiometric model of a given physiology (Figure 3.1). We follow this workflow to discuss our results. At first, we identify the bi-directional reactions and determine feasible flux directionality profiles (FDPs). In a FDP, reactions can only operate in a unique direction. We discuss how alternative FDPs affect the conclusions of kinetic models. We then consider how the flux values and the metabolite concentration levels within a FDP affect kinetic model predictions. The MCA outputs of the kinetic models are studied to systematically derive metabolic engineering decisions. For further information on the methodologies used, we refer the reader to the methods section of the chapter.

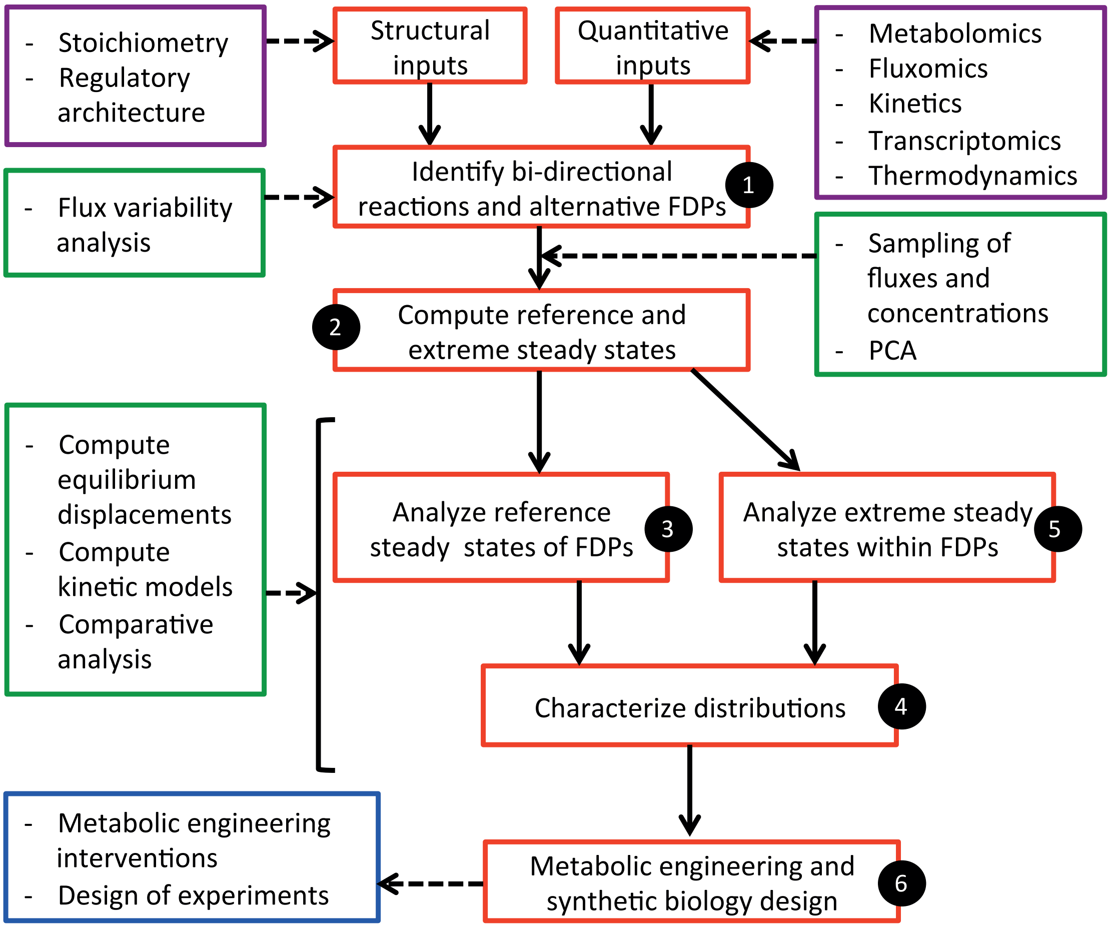

- Stoichiometry
- Regulatory architecture
Structural inputs
Quantitative inputs
- Metabolomics
- Fluxomics
- Kinetics
- Transcriptomics
- Thermodynamics
- Flux variability analysis
Identify bi-directional reactions and alternative FDPs
1
- Sampling of fluxes and concentrations
- PCA
2
Compute reference and extreme steady states
- Compute equilibrium displacements
- Compute kinetic models
- Comparative analysis
Analyze reference steady states of FDPs
3
Analyze extreme steady states within FDPs
5
Characterize distributions
4
- Metabolic engineering interventions
- Design of experiments
Metabolic engineering and synthetic biology design
6

Figure 3. 1. Procedure for characterizing and analyzing multiplicities in metabolic networks. The procedure consists of several computational steps wherein the available data are integrated, the alternative solutions are identified, the populations of non-linear models are built, and the output variables are analyzed to make robust conclusions (for details see main text).

3.2.1. Multiplicity of flux directionality profiles

To derive a reduced *E. coli* metabolic model from the iJO1366 GEM [77], we used the redGEM and lumpGEM algorithms as they provide a systematic and modular way for reducing GEMs, whilst preserving growth and gene essentiality [57, 58]. The obtained core stoichiometric model of the *E. coli* metabolism consisted of 277 reactions and 160 metabolites distributed over the cytosol and the extracellular space (Methods). To constrain the model and derive alternative steady states, we integrated fluxomics and metabolomics data, as we did in chapter two [61] (Appendix A1) within the thermodynamic formulation (Appendix B1) of the stoichiometric model for aerobically grown *E.coli*. Hence, we set the glucose uptake to 7.54 mmol/gDW/h, the growth rate to 0.61 /h and, the excretions of acetate, formate and succinate to 3.5 mmol/gDW/h, 0.5 mmol/gDW/h and 10^{-4} mmol/gDW/h, respectively (Figure 3.2A). For the reactions previously reported in literature we used assumptions about reaction directionalities [61, 79-82] (Methods). Thermodynamic-based variability analysis (TVA) [83] suggested the presence of seven bi-directional reactions in our model: fumarase (FUM), triose-phosphate isomerase (TPI), ribulose-5-phosphate 3-epimerase (RPE), transaldolase (TALA), transketolase 1 (TKT1), transketolase 2 (TKT2), and glucose-6-phosphate isomerase (PGI). All combinations of these seven reactions operating in one or the opposite direction could theoretically lead to up to 128 (2^7) FDPs. However, due to the stoichiometric and thermodynamic coupling in the network, only 25 out of 128 FDPs were feasible.

Some of these reactions such as PGI and FUM are commonly considered as unidirectional. However, Rabinowitz and coworkers reported that these seven identified reactions are bi-directional in *E.coli*, yeast and immortalized baby mouse kidney cells [62]. This suggests that, for previously uncharacterized physiologies and/or for reactions with no fluxomics data, we should consider all feasible reaction directionalities. This is a way of ensuring that we account for the flexibility of cellular metabolism. For simplicity and clarity of further discussion, we wanted to analyze four FDPs with the most distinct physiologies out of 25 feasible ones. We assumed that changing the directionality of reactions with the largest TVA flux range would result in the most distinct FDPs. PGI and FUM had the largest feasible TVA flux ranges from the seven bi-directional reactions. Hence, to generate these four distinct FDPs (Figure 3.2B), we were changing the directionality of both PGI and FUM in either the forward or backward direction (Figure 3.2C) while keeping the directionalities fixed for the 5 remaining bi-directional reactions. The directionality of these 5 bi-directional reactions (other than PGI and FUM) was determined as follows. We first defined and calculated the flux variability score for each of the 25 FDPs (Methods). A higher flux variability score suggests that the reactions of the FDP are on average more flexible and can operate in relatively wider flux ranges. We then took the directionalities of the remaining five bi-directional reactions from the FDP with the highest score. In this study, we assessed the model predictions and their implications on metabolic engineering decisions around these four FDPs. Nevertheless, different study-dependent criteria for selecting the FDPs could be devised.

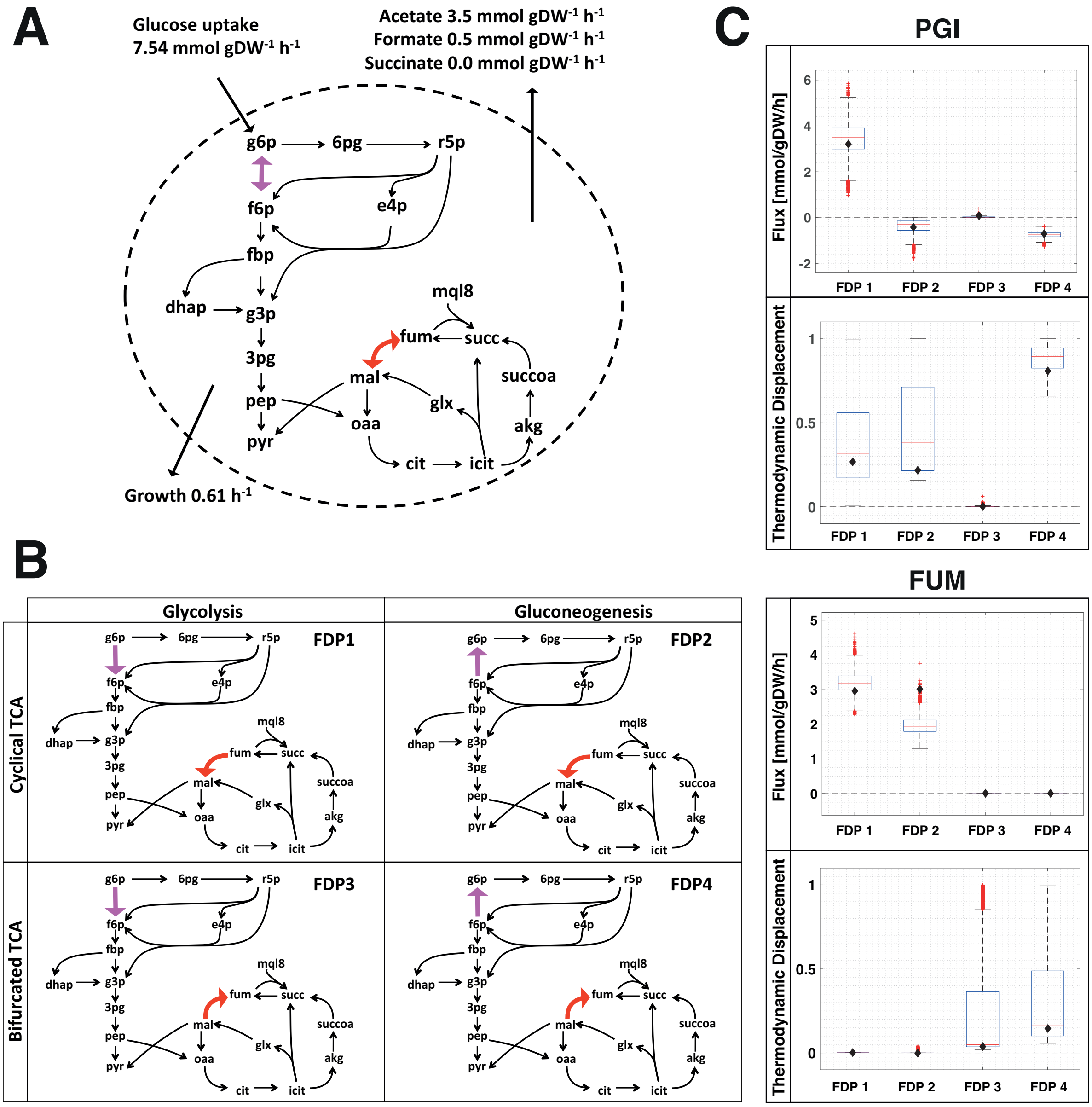


Figure 3. 2. Multiple operational configurations for the same observed physiology of aerobically grown *E. coli*. (A) Representation of *E. coli* network. The fluxomics data that were integrated are indicated as uptake, secretion and growth rates. The bidirectional reactions are colored: phosphoglucose isomerase, (PGI, magenta) and fumarase, (FUM, red). (B) Representation of the four FDPs for the physiology under study. (C) Flux and thermodynamic displacement distributions of PGI and FUM reactions for each of the four generated FDPs. The boxplots show distributions for 5,000 samples. The central red line indicates the median, and the bottom and top edges of the box indicate the 25th and 75th percentiles, respectively. The whiskers correspond to approximately ± 2.7, which is the standard deviation, or 99.3% coverage if the data are normally distributed. Outliers are the

points not covered by the range of the whiskers and are plotted individually using the '+' symbol. The black diamond is the RSS value. Full metabolite names are given in supplementary materials (Appendix B1).

3.2.2. Comparative analysis of alternative flux directionality profiles

3.2.2.1. Reference steady states (RSSs) of FDPs

In building kinetic models, we must have steady state flux values and metabolite concentrations around which we construct them. In the case of uncertainty, we sampled steady states for the flux values and the metabolite concentrations for each FDP and used principal component analysis (PCA) to select their RSSs (Methods). There were considerable differences in the RSS values for the fluxes and thermodynamic displacements of reactions across the network, particularly for TPI, enolase (ENO), phosphogluconate dehydrogenase (GND), and aconitase A (ACONTa) in the central carbon metabolism (Figure 3.3). This is because the relative activity of the oxidative tricarboxylic acid (TCA) cycle, the glyoxylate shunt, and both the oxidative and the non-oxidative pentose phosphate pathway (PPP) change between FDPs. Since PGI and FUM are the only two reactions changing directionalities amongst the four FDPs, it is reasonable to expect the most affected fluxes of reactions to be in their topological vicinity, which is true for GND, TPI, ACONTa, and succinate dehydrogenase (SUCDi) (Figure 3.3). However, we found large changes in flux magnitudes across the FDPs that were associated with reactions farther away from FUM and PGI, such as the electron transport chain (ETC) reactions, NADH dehydrogenase (NADH16pp) and NAD transhydrogenase (NADTRHD). The TVA studies explain this as the ETC compensates in FDPs 2-4 for producing NADPH (Appendix B2 and B3). Additionally, the RSS flux value for GND was considerably smaller in FDP1 than in the other FDPs, resulting in reduced NADPH production via the oxidative branch of the PPP that is coupled with the ETC (Appendix B2). For further comparative TVA studies of the FDPs, we refer the reader to supporting information (Appendix B3).

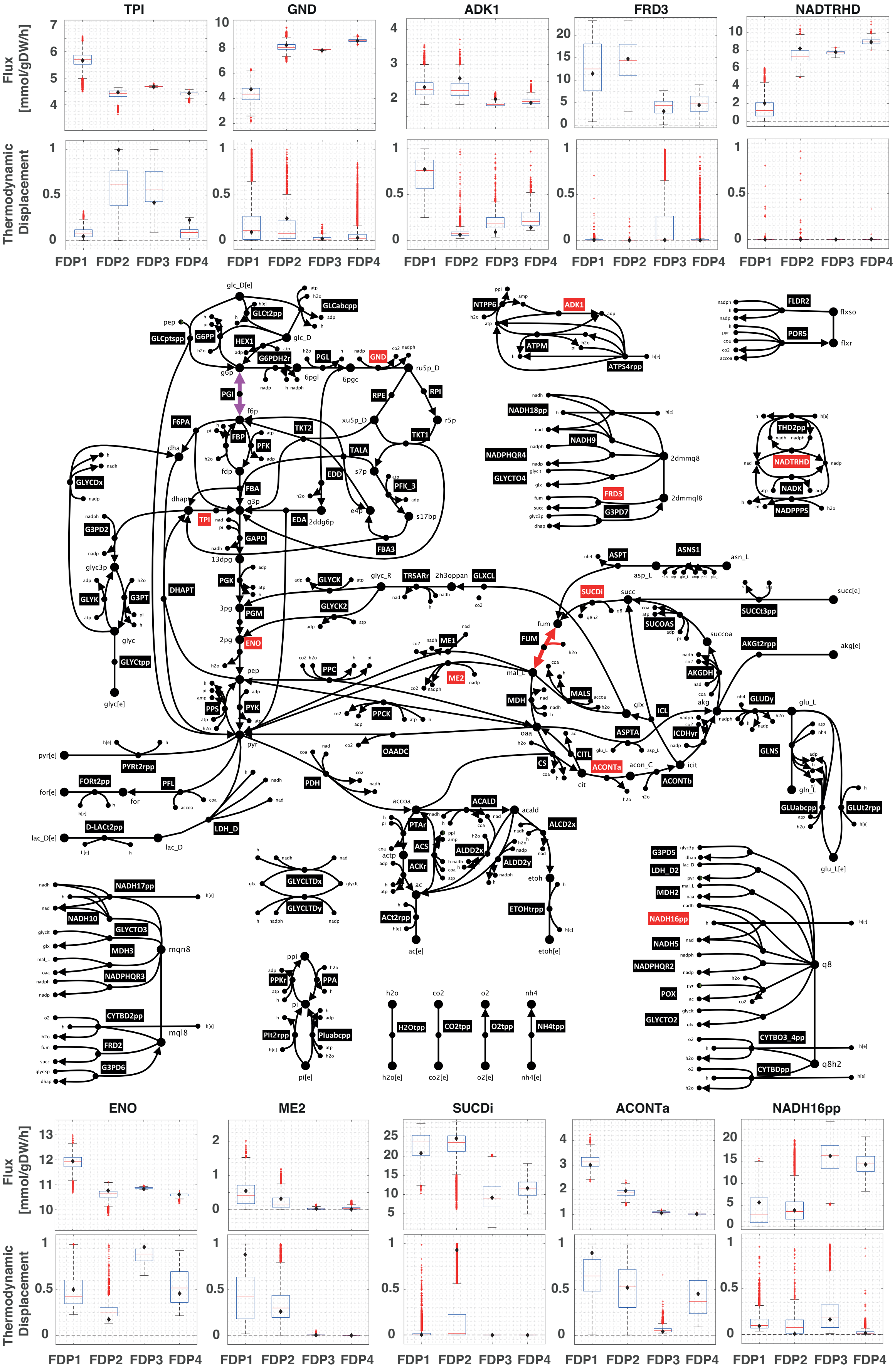

TPI
GND
ADK1
FRD3
NADTRHD
Flux [mmol/gDW/h]
Thermodynamic Displacement
FDP1 FDP2 FDP3 FDP4
glc_D[e]
GLCt2pp
GLCabcpp
GLCptspp
G6PP
HEX1
G6PDH2r
PGL
GND
glc_D
g6p
6pgl
6pgc
ru5p_D
PGI
RPE
RPI
f6p
xu5p_D
r5p
F6PA
TKT2
TKT1
FBP
PFK
TALA
dha
fdp
s7p
EDD
PFK_3
GLYCDx
FBA
g3p
e4p
dhap
TPI
EDA
2ddg6p
s17bp
G3PD2
GAPD
FBA3
13dpg
glyc3p
DHAPT
PGK
GLYCK
glyc_R
TRSARr
2h3oppan
GLXCL
G3PT
GLYK
3pg
PGM
GLYCK2
glyc
2pg
ENO
GLYCtpp
glyc[e]
pep
PPC
ME1
ME2
PPS
PYK
PPCK
pyr
OAADC
pyr[e]
PYRt2rpp
FORt2pp
PFL
for[e]
for
PDH
D-LACt2pp
LDH_D
lac_D[e]
lac_D
NTPP6
ADK1
ATPM
ATPS4rpp
FLDR2
POR5
flxso
flxr
NADH18pp
NADH9
NADPHQR4
GLYCTO4
FRD3
G3PD7
2dmmq8
2dmmql8
THD2pp
NADTRHD
NADK
NADPPPS
ASPT
ASNS1
asn_L
asp_L
SUCDi
succ
succ[e]
SUCCt3pp
SUCOAS
succoa
fum
FUM
AKGt2rpp
akg[e]
AKGDH
mal_L
MALS
MDH
glx
ICL
akg
GLUDy
glu_L
oaa
ASPTA
ICDHyr
CITL
ACONTa
acon_C
icit
ACONTb
CS
cit
GLNS
gln_L
GLUabcpp
GLUt2rpp
glu_L[e]
accoa
ACALD
acald
PTAr
ALCD2x
actp
ACS
ALDD2x
ACKr
ALDD2y
etoh
ac
ETOHtrpp
ACt2rpp
ac[e]
etoh[e]
G3PD5
LDH_D2
MDH2
NADH16pp
NADH5
NADPHQR2
POX
GLYCTO2
q8
CYTBO3_4pp
CYTBDpp
q8h2
NADH17pp
NADH10
GLYCTO3
MDH3
NADPHQR3
mqn8
CYTBD2pp
FRD2
G3PD6
mql8
GLYCLTDx
GLYCLTDy
ppi
PPKr
PPA
pi
PIt2rpp
PIuabcpp
pi[e]
h2o
co2
o2
nh4
H2Otpp
CO2tpp
O2tpp
NH4tpp
h2o[e]
co2[e]
o2[e]
nh4[e]
ENO
ME2
SUCDi
ACONTa
NADH16pp
Flux [mmol/gDW/h]
Thermodynamic Displacement
FDP1 FDP2 FDP3 FDP4

Figure 3. 3. Optimally grown aerobic *E. coli* metabolic network. Each of the 10 reactions labeled in red has an associated graph with the respective flux and thermodynamic displacement distributions for each FDP. The boxplots show distributions for 5,000 samples. The central red line indicates the median, and the bottom and top edges of the box indicate the 25th and 75th percentiles, respectively. The whiskers correspond to approximately ± 2.7, the standard deviation, or 99.3% coverage if the data are normally distributed. Outliers are the points not covered by the range of the whiskers and are plotted individually using the '+' symbol. The black diamond is the RSS value. Full enzyme and metabolite names are given in supplementary materials (Appendix B1).

The differences in the RSS concentration vectors across the FDPs translate into distinctive distributions of the Gibbs free energy across the networks for each FDP. The metabolite concentration values in the RSSs varied the most across the FDPs for the reaction cofactors NAD^+, NADH, NADP, AMP, and ATP (Appendix B2). We also noticed significant differences in some central carbon metabolite RSS concentrations, such as: 6-phospho-D-gluconate, D-glucose-6-phosphate, D-fructose-6-phophate, D-xylulose 5-phosphate, sedoheptulose 7-phosphate, D-erythrose 4-phosphate, phosphoenolpyruvate, fumarate, L-malate, citrate, and oxaloacetate (Appendix B2). These metabolites and the aforementioned cofactors participate in most of the network reactions, causing the thermodynamic displacements of reactions including GND, NADH16pp, SUCDi, adenylate kinase (ADK1), and ME2 to change considerably across RSSs of the FDPs (Figure 3.3). As observed for the RSS fluxes, reactions that were either topologically close to the bidirectional reactions FUM and PGI and some topologically distant reactions in the ETC displayed the most considerable changes in thermodynamic displacement (Figures 3.2 and 3.3). It is particularly important to recognize that the change in directionality of one reaction between two FDPs can actually cause thermodynamic displacements to change across the whole metabolic network (Figure 3.4). Hence, the FDP affects the Gibbs free energy distribution across the network, which in turn can affect a greater part of the network than just the topological neighbors of the bi-directional reactions that change directionality between FDPs.

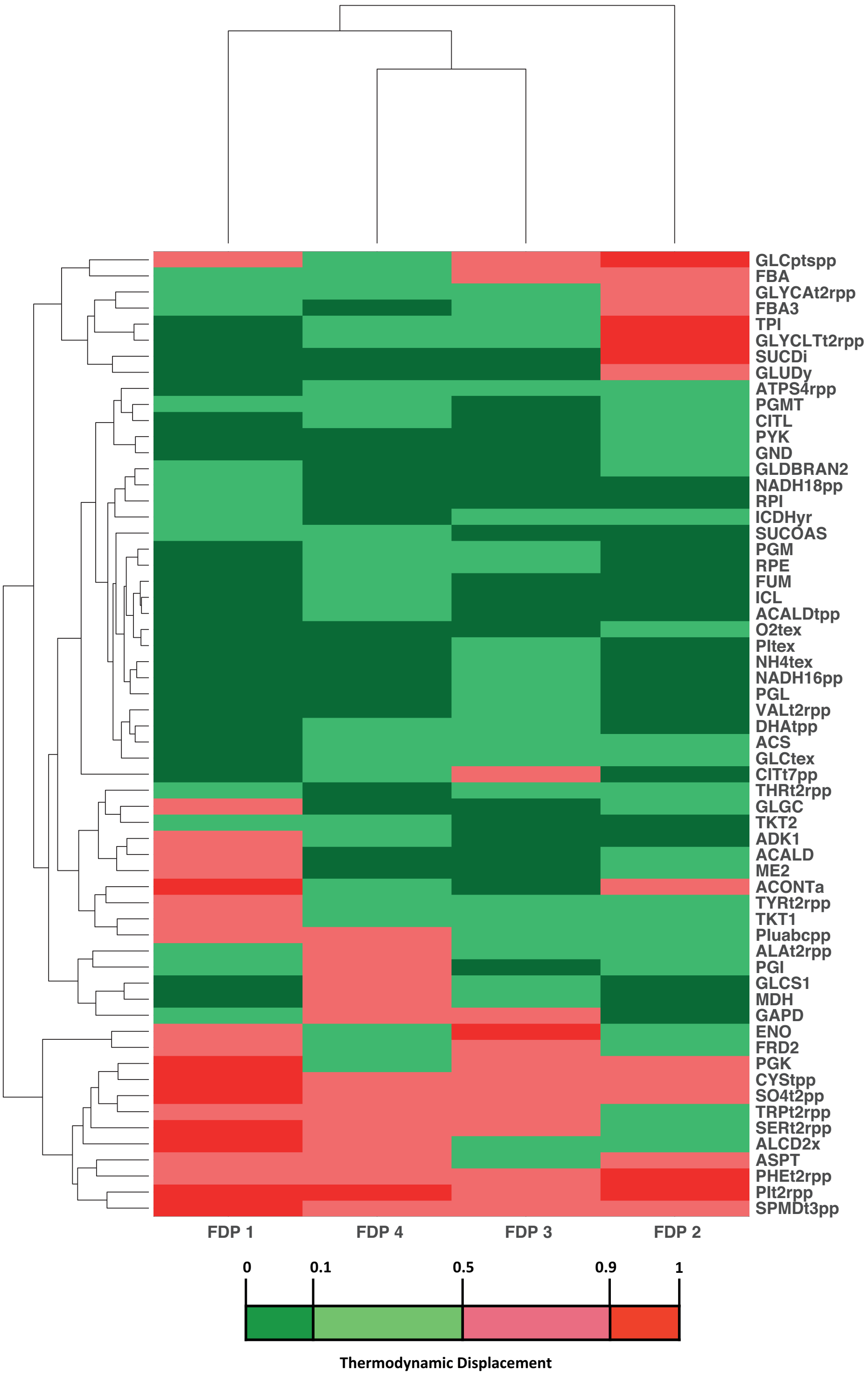
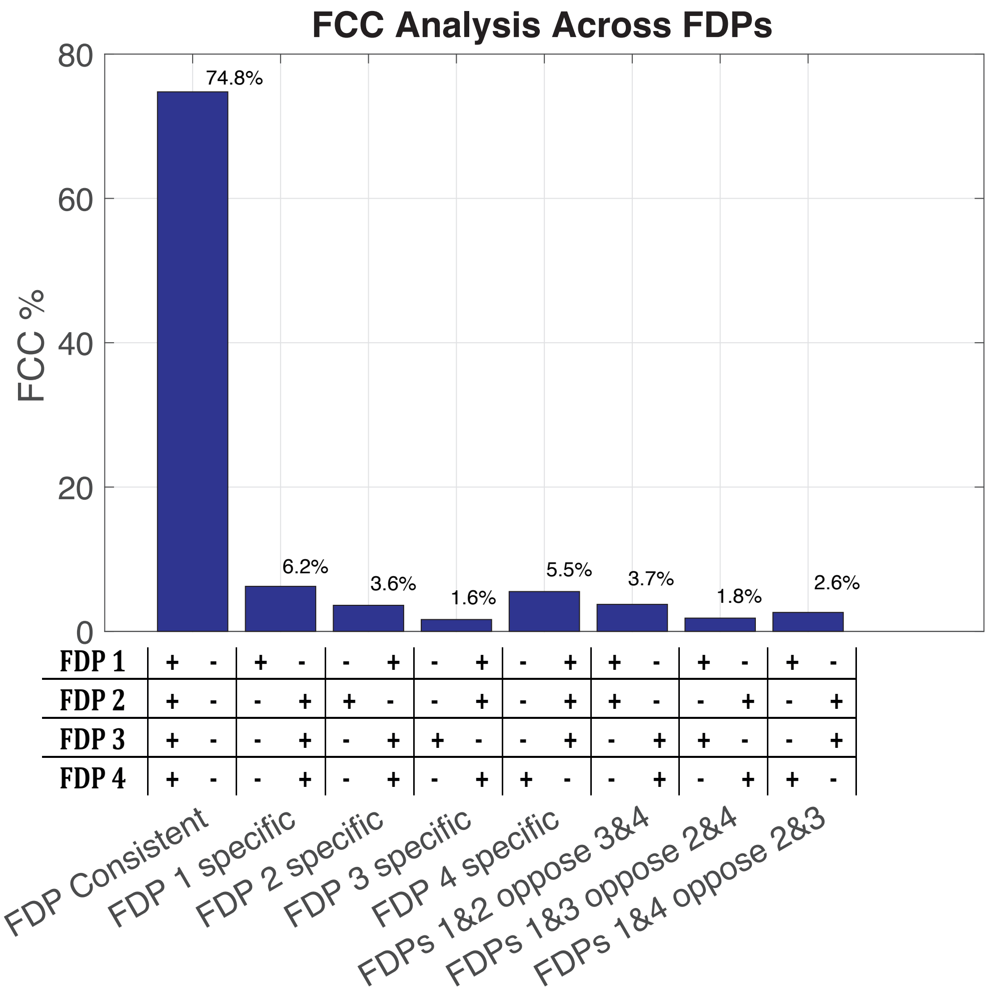


Figure 3. 4. Cluster analysis of thermodynamic displacements across FDPs. Heat map showing thermodynamic displacements of the reactions across each FDP. Reactions differed the most in thermodynamic displacement across FDPs based on categorization of displacements (Methods). The rows represent the similarity between reactions and the columns represent the similarity between FDPs. The distances between the dendrograms were computed based on the Euclidean distance between the thermodynamic displacements, both column- and row-wise. Full enzyme names are given in supplementary materials (Appendix B1).

3.2.2.2. Analysis of control patterns

Controlling the levels of various enzymes in a target organism can help to achieve the desired levels of bioengineered products or metabolites. Determining the degree of control of various enzymes in each FDP can help find the key locations to target, and we did this by sampling the kinetic parameter space uniformly (Methods) using the ORACLE [48, 50, 51, 84-86, 98] workflow, generating a population of 50,000 stable kinetic models for each FDP. The kinetic parameter space was sampled based on the degree of saturation of the enzyme active sites, as proposed previously by Hatzimanikatis and colleagues [84]. ORACLE verifies the local stability of the model around the steady state by testing that the Jacobian matrix has no positive real eigenvalues for the sampled set of parameters. We then calculated, for the stable models, the flux control coefficients (FCCs), representing the fold change in a specific flux with respect to the perturbation of an enzyme's activity, of 275 enzymatic reactions with respect to their enzymes. We then compared the differences in FCCs across FDPs for the populations of stable kinetic models.

If the signs of a FCC are not the same across FDPs, the FCC depends on the FDP, and making metabolic engineering decisions is ambiguous. This means that the alternative steady states have a significant impact on the FCC, and we should be careful when deriving conclusions. FCC values with an absolute mean value larger than 0.1 across all the FDPs have significant control over the fluxes in the network (Methods). Fluxes smaller than 0.01 mmol/gDW/h were not considered, as we focused around central carbon metabolism. To investigate the differences in control patterns for each FDP, we compared the sign of these FCCs across the FDPs (Figure 3.5) because the sign determines the increase or decrease in magnitude of a flux upon perturbation of an enzyme level. Hence, the sign can indicate if it may be possible to overexpress, down-regulate, or even suppress a gene to achieve a target enzyme level for bioengineering purposes. If the signs of the mean FCCs are equal across all FDPs, we have consensus, and the FCCs are independent of the FDP. This indicates that the predictive conclusions drawn should be valid for all the tested alternative steady states, suggesting that our metabolic engineering conclusions are more robust. Nearly 75% of the FCCs studied agreed in sign across all the FDPs, meaning that most metabolic flux response predictions are consistent (Figure 3.5), though the 25% of potentially inconsistent predictions highlights the importance of considering alternative steady states. As we sampled the kinetic parameter space uniformly for the FDPs, differences in the thermodynamic displacement (Figure 3.4) between these FDPs are the main reason behind these variations in their control pattern. Further discussions around these differences are in the supplementary document (Appendix B3).



FDP 1	+	-	+	-	-	+	-	+	-	+	+	-	+	-	+	-
FDP 2	+	-	-	+	+	-	-	+	-	+	+	-	-	+	-	+
FDP 3	+	-	-	+	-	+	+	-	-	+	-	+	+	-	-	+
FDP 4	+	-	-	+	-	+	-	+	+	-	-	+	-	+	+	-

Figure 3. 5. General statistics on FCCs across FDPs. Histogram displaying the fraction of reactions that have FCCs with a certain sign pattern across the FDPs. There are three main categories of the FCCs: (i) consistent among all FDPs, (ii) FDP specific, and (iii) two FDPs contradicting two other FDPs. The FCCs were averaged over the 50,000 samples for each FDP, and the ones selected for analysis had a mean absolute value larger than 0.1 (10% fold change). For example, to assist reading the figure, the column "FDP 1 specific" has two possible scenarios as it contains FCCs that are positive in FDP1 and negative in the other three FDPs as well as FCCs that are negative in FDP1 and positive in the three other FDPs.

3.2.2.3. Ranking target enzymes for flux control

Some of the fundamental biological tasks performed by a cell include substrate uptake, product excretion, and cellular growth, represented by μ. Since we modeled the physiology of optimally grown *E. coli*, we first studied what enzymes have control over cellular growth. These enzymes were considered as attractive target candidates for genetic manipulation to improve cellular growth. We selected the top five enzymes with high absolute FCC values for each FDP and computed the control exerted by these enzymes over cellular growth (Figure 3.6). Several enzymes such as PGM, RPE, TPI, PPC, and NAD kinase (NADK) had considerably different control patterns for μ across FDPs in terms of magnitude and sign. Because of the abovementioned differences in the thermodynamic displacement of enzymes across the FDPs, it was not surprising to see opposing FCC signs. For instance, TPI is far away from equilibrium for FDP1 and near equilibrium for FDP2 (Figure 3.4), resulting in different conclusions when considering control coefficients of cellular growth (Figure 3.6). In contrast, PGM is always far away from equilibrium but, due to kinetic coupling, has considerably different degrees of control across FDPs, indicating the importance of considering alternative steady states. More importantly, we also found enzymes that agreed in terms of sign across the FDPs. NADTRHD, phosphofructokinase (PFK), and ATP maintenance (ATPM) were the top target enzymes – independent of the FDP – for improving the cellular growth of optimally grown *E. coli*.

Because in the studied physiology, growth is based solely on glucose, we decided to study how consistent the FCCs of glucose uptake via D-glucose transport (GLCptspp) were across FDPs. PGM, PFK, RPI, and RPE agreed across all the FDPs in terms of sign and the magnitude of their FCCs, making them attractive metabolic engineering targets for increasing GLCptspp flux (Appendix B3, Figure S4). Based on a consistent magnitude across all the FDPs, PGM and PFK were the top two target enzymes that seemed to control the glucose uptake of optimally grown *E. coli*. TPI, PPC, NADTRHD, glucose 6-phosphate dehydrogenase (G6PDH2r), and 6-phosphogluconolactonase (PGL) control GLCptspp in at least one FDP but not across all. As for the control of cellular growth, the differences in thermodynamic equilibrium and kinetic coupling between the FDPs explain these results.

These observations emphasize the importance of considering enzyme kinetics and the existence of alternative steady states before making metabolic engineering conclusions based off kinetic models, especially given that further similar observations were made for FCCs of other fluxes. Generally, we noticed that the enzymes whose control remained unchanged across FDPs were found in the central carbon metabolism. Inconsistent enzyme control was observed in peripheral and transport reaction enzymes topologically further away from the central carbon.

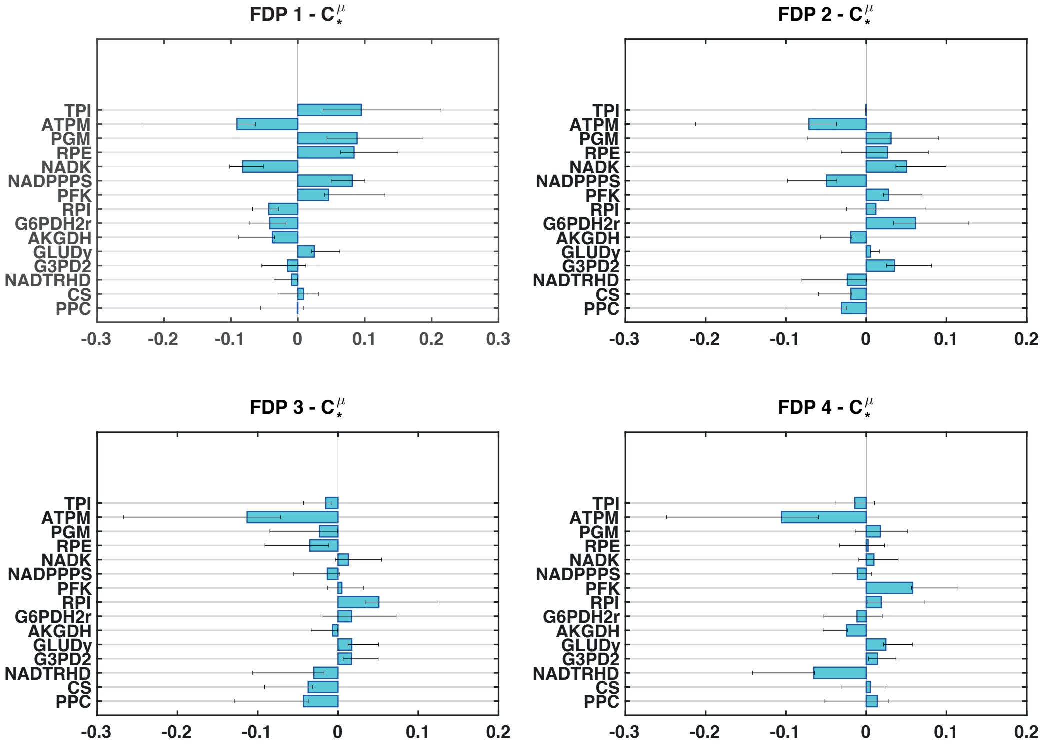


Figure 3. 6. Flux control across FDPs for cellular growth, μ. Illustration of the union of the top five enzymes across the FDPs in terms of absolute control over cellular growth. The whiskers correspond to the upper and lower quartiles of the 50,000 FCC populations, and the bars correspond to the means. Full enzyme names are given in supplementary materials (Appendix B1).

3.2.2.4. Study of uncertainty in FCCs

To characterize the variability of cellular growth FCCs with respect to all central carbon enzymes (i.e. no transporter nor exporter) for the populations of 50,000 models (Figure 3.7), we studied the uncertainty across the four FDPs using PCA (Methods). The first two principal components (PCs) covered a majority of the variance, with 93%, 62%, 56%, and 69% for FDP1–4, respectively. For FDP2–4, at least seven, eight and six PCs, respectively, were required to account for more than 90% of the variance between the FCC populations. This suggests that the uncertainty in the cellular growth FCCs was considerably more distributed for FDP2–4 than for FDP1 that required only two.

In PCA, each variable has a score on the PCs that are under consideration, which corresponds to its contribution to the variability described by the given PC. In FDP1, the growth FCC of the enzymes NADK and NADPPPS corresponded to the highest PC scores above absolute 0.5 on the first PC, suggesting that most of the uncertainty comes from these ETC enzymes (Figure 3.7). Their scores on both PCs were strongly opposed in terms of sign, suggesting that these FCCs anti-correlate. In fact, the cellular growth FCCs of NADK and NADP phosphatase (NADPPPS) had a -1.00 Pearson correlation coefficient, further indicating that they were exactly anti-correlated. On the other hand, enzymes PGM and RPM had very similar PC scores, and we note that the correlation coefficient of these FCCs was 0.91, indicating a near-perfect correlation.

Similarly to FDP1, we studied FDP2–4 to find underlying covariance patterns between the cellular growth FCCs (Figure 3.7). We noticed that certain trends were preserved between the FDPs as, for instance, the PCA scores of NADPPPS and G6PDH2r tended to have an opposing sign across the four FDPs for at least one of the plotted PCs. In fact, NADPPPS and G6PDH2r cellular growth FCCs had Pearson correlation coefficients of -0.81, -0.84, -0.53, and -0.71 for FDP1–4, respectively. Hence, we can use PCA to explore and unravel covariance patterns in FCCs to understand their underlying functional relationships. Although fully describing the relationship between FCCs remains a non-trivial task, PCA makes strides towards interpreting the various sources of uncertainty.

NADTRHD, PFK, and ATPM were the top candidates for improving cellular growth, as determined previously based on absolute means (Figure 3.6). If we had to select one of these three enzymes for genetic engineering, we want it to be the one with the least uncertainty for this purpose. We observed that PFK scores lower than NADTRHD and ATPM on the PCs across all the FDPs (Figure 3.7), demonstrating the least uncertainty and suggesting that it could be the most prominent target enzyme. A similar analysis could be performed for other FCCs, such as glucose uptake (Appendix B3, Figure S5). We conclude that to improve growth of aerobically grown *E. coli,* PFK would also be a top candidate enzyme to metabolically engineer, despite the uncertainties.

The uncertainty in the kinetic parameters and its impact on our studies remains difficult to quantify due to the underdetermined nature of this highly non-linear solution space but it has been further characterized by previous work done by Andreozzi *et al.* [49]. For our study, when we next compare the effect of uncertainties stemming from the flux and the concentration steady-state solutions, we decided to fix the distributions of the sampled enzyme saturations using beta distributions (Methods). This allowed us to keep the same level of uncertainty in all the kinetic parameters for our comparisons.

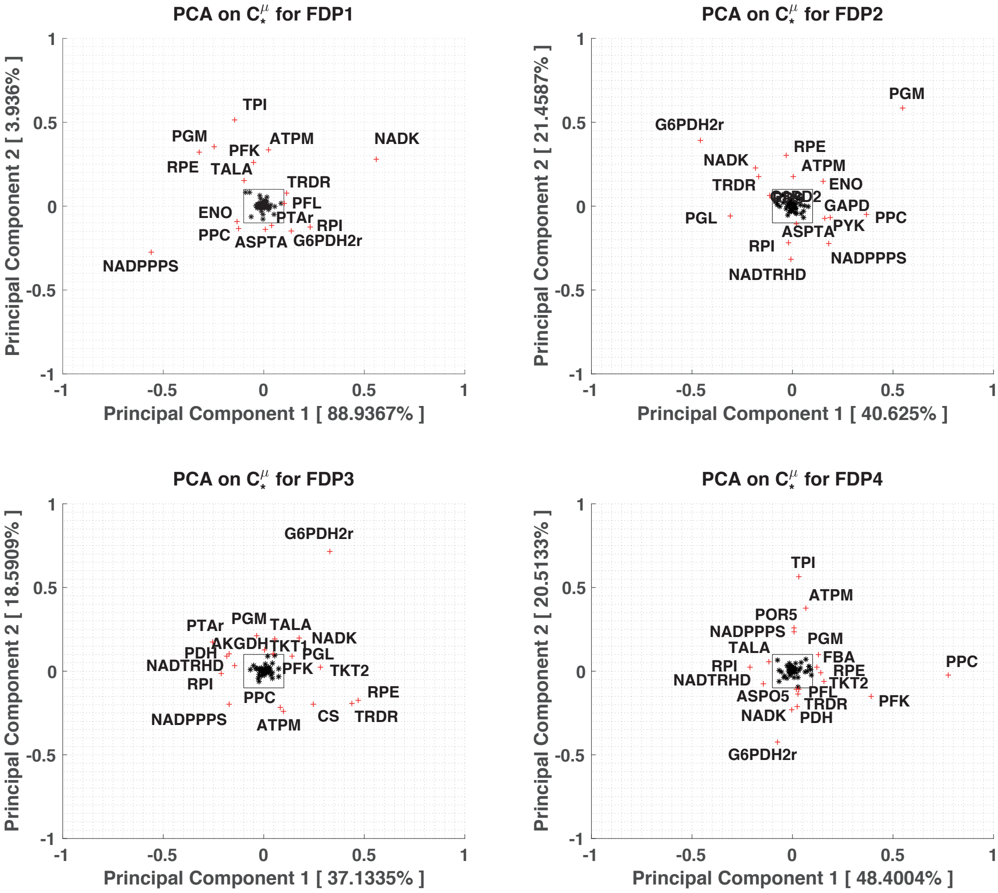


Figure 3. 7. Principal component analysis (PCA) of FCCs of cellular growth across FDPs. PCA was carried out on the cellular growth FCCs of the 50,000 samples for each FDP. Only cellular growth FCCs with respect to non-transporter and non-exporter reactions were considered. The two principal components, namely PC1 and PC2, were plotted to study the variance in the FCC samples. The values in brackets correspond to the variance covered by the principal components (PCs). Full enzyme names are given in supplementary materials (Appendix B1).

3.2.3. Impact of flux and concentration profiles

We next assume that we know the directionality of each reaction in our network but the system still remains underdetermined and we have multiple feasible flux and concentration steady states within the allowable solution space. Because of this, we then studied how the underlying uncertainty that results in alternative steady states within a FDP affected the predictions of kinetic models. For this analysis, we selected FDP1 because it has: (1) reaction directionalities corresponding to the more frequently observed *E. coli* wildtype operational state of glycolysis and TCA cycle [62, 106-108], (2) the largest flux variability score, and (3) the highest specificity in control (Figure 3.5). FDP1 was then more exhaustively sampled with 100,000 iterations of concentrations and fluxes. We chose RSSs from the flux and concentration samples as previously done with the four FDPs and also used PCA to determine four extreme steady states (ESSs) for the concentrations and four ESSs for the flux solution spaces (Methods). The ESSs are samples with the most distant behavior from the "average" displayed by the RSS. An ESS is a steady state that is located along a PC at one of its extremes and can be used to characterize the "extreme" behaviors of the FDP1.

To study the impact of flux and concentration ESSs on MCA outputs, we had to decouple their effects (Methods), so we isolated the effects of flux and concentration separately in our analysis. Therefore, when we studied the effect of flux, we kept the same concentration RSS and paired it with the four flux ESSs, meaning that we had four pairs of flux ESSs with the same RSS concentration. Similarly, when we decoupled the effect of concentration, we paired the flux RSS with the four concentration ESSs. Therefore, we had a total of eight extreme pairs of flux and concentration steady states to study. We compared these extreme pairs to the reference case, where we had the flux and concentration RSSs paired. For the reference case, we sampled the saturation state space for 50,000 stable kinetic models. We used the distributions of the kinetic parameters from this reference case to generate models for the ESSs (Methods). We sampled 50,000 stable kinetic models for each ESS and computed the FCCs using MCA. We performed a comparative analysis like the comparative analysis of the FDPs to assess the degree of confidence of our conclusions with respect to both the extreme flux profiles and the extreme concentration profiles.

3.2.3.1. Flux uncertainty propagation to control

Like the comparison of FDPs, the ESS flux profile magnitudes mainly differed in peripheral fluxes, such as glutamate transport, glycogen metabolism, and ETC reactions (Appendix B2). Noticeable differences in the central carbon fluxes of greater than 1 mmol/gDW/h were seen in pyruvate kinase (PYK), fumarate reductase (FRD3), ME2, NADH17pp, NADH18pp, and NADTRHD. To assess how this variability in fluxes affected the degree of confidence in our MCA conclusions, we considered control over glucose uptake and cellular growth. For glucose uptake, the top enzymes of the flux ESSs, aspartate transaminase (ASPTA), PFK, AKGDH, TKT1, ENO, and TPI, are reasonable candidates for improving uptake because they are all qualitatively in agreement across the flux ESSs and have a control value larger than absolute 0.1, indicating significant control over their networks. This excludes PDH (Figure 3.8A). Citrate synthase (CS) and G6PDH2r, may appear to be attractive targets based on some of the ESSs, but since this property is not in consensus agreement across all ESSs, they are less reliable targets. The top enzymes controlling cellular growth were sensitive to the ESSs, as just 47% of them were in agreement sign-wise with each other (Appendix B3, Figure S6). However, we can still find reliable target enzymes, such as ENO, AKGDH, and glutamate dehydrogenase (GLUDy), mainly in the central carbon reactions that have a reasonable magnitude and consensus agreement.

3.2.3.2. Concentration uncertainty propagation to control

We repeated the previous analysis for ESS concentration vectors to study how they impact MCA outputs. The main concentration differences between the extreme concentration vectors were for amino acids (R-glycerate, L-glutamine, L-lysine, D-alanine, L-proline), inorganics (potassium, iron, and cobalt), cofactors (NAD and AMP), and several biomass building blocks (Appendix B2). We considered the top enzymes for glucose uptake FCCs (Figure 3.8B) and noticed that the MCA conclusions were more sensitive to concentration values than to variations in flux values. Candidate target enzymes to improve glucose uptake would be PFK, GAPD, ENO, and RPI, as at least three out of five of the steady states are consistent and have a control value larger than 0.1 (Figure 3.8B). Hence, the metabolic engineering decisions derived from the MCA outputs appear to be more sensitive to concentration values rather than flux values. As the biomass building block and amino acid metabolite concentrations were changing between these ESS concentration vectors, it makes sense that they would have a higher FCC variability. The concentration values in turn directly impact the thermodynamic displacements and the enzyme saturation states, which impact the MCA conclusions. The cellular growth FCCs were very sensitive to the ESS metabolite concentration vectors because most them were in sign disagreement (Appendix B3, Figure S6). NADTRHD and ATPM were the most appealing enzymes for controlling cellular growth due to sign and magnitude consistency across the ESSs of their FCCs.

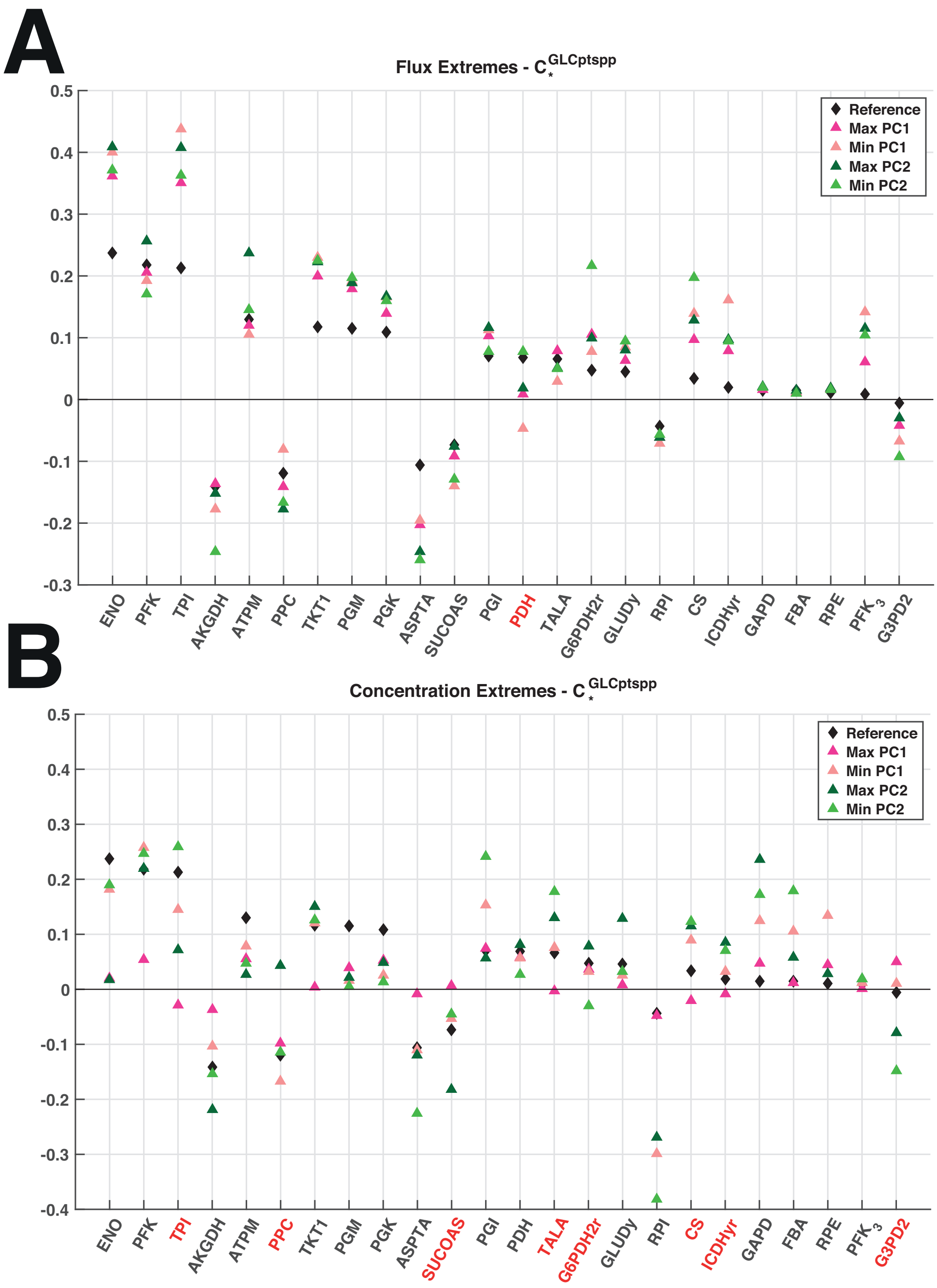


Figure 3. 8. Flux control patterns across extreme steady-state solutions. Illustration of the union between the top 10 enzymes across the (A) flux and the (B) concentration ESSs in terms of absolute control over glucose uptake. FCCs were sorted in decreasing order of absolute magnitude of the RSSs (reference). The enzyme names in black indicate that the FCCs were sign-consistent (in agreement) and red if they were sign-inconsistent (in opposition). Full enzyme names are given in supplementary materials (Appendix B1).

3.3. Conclusion

This chapter studied the impact of alternative concentration and flux steady states on the conclusions derived from the MCA outputs of the non-linear kinetic models built around them using the physiology of optimally grown *E. coli*. We show that different FDPs can lead to distinct metabolic engineering conclusions when analyzing output FCCs of the non-linear models. The ME2, PPC, and PGI examples illustrate how thermodynamics and kinetic coupling can change the control from one FDP to another. These enzymatic reactions were topologically close to the bidirectional reactions that changed between the FDPs, though, less intuitively, we noticed that there were changes in thermodynamic displacements across FDPs in enzymes that were topologically far away from the bidirectional reactions. We then studied the uncertainty within a single FDP, and using PCA to study the extremes of the solution space, found that within a FDP, MCA outputs appeared to be more sensitive to concentration values rather than flux values. These observations emphasized the importance of considering alternative solutions when studying a physiology as the steady state affects directly the decisions for hypothesis generation in basic research and design in synthetic biology and metabolic engineering. Hence, we propose a workflow for assessing this uncertainty to make more reliable metabolic engineering decisions that can be broadly applied to any kinetic model to improve the predictions resulting from it.

We then used our workflow to pick target enzymes for genetic modification, identifying NADTRHD, PFK, and ATPM as the top target enzymes independent of the FDP for improving the cellular growth of optimally grown E. coli. PFK and PGM were selected as top enzymes independent of the FDP for improving glucose uptake of optimally grown E. coli. We stress the importance of selecting target enzymes that exhibit control across all the FDPs to make more reliable decisions, highlighting the need to consider alternative steady states when building non-linear kinetic models for a given physiology, as they have imminent implications on the conclusions derived from the MCA. The herein proposed workflow can be used to suggest metabolic engineering decisions for a given study and can provide insights into the design of experiments, as the ranking of candidate enzymes can highlight reactions or enzymes that need further characterization and study due to their variability. In this light, the next chapter discusses how statistical methods can be used to attribute confidence intervals to variables of kinetic models and their outputs.

3.4. Materials and methods

3.4.1. Reduced *E. coli* model

The model stoichiometry for this study was derived from *E. coli* iJO1366 [77] using redGEM, a systematic framework for developing core models that are consistent with their genome-scale counterparts [57, 58]. The resulting reduced models are context-specific and in the process of reduction it is important to define the carbon sources, the content of media and also the metabolic subsystems of interest for the study. We used a minimal media with glucose as the sole carbon source and the selected starting subsystems were ones pertaining to central carbon metabolism (glycolysis/gluconeogenesis, citric acid cycle, pentose phosphate pathway, pyruvate metabolism, and glyoxylate metabolism). Omics data for the physiology of optimally grown *E. coli* under aerobic conditions were extracted (Appendix A1) from McClosekey *et al.*[61]. The data were integrated in the form of constraints into the MILP formulation of the thermodynamics-based metabolic flux analysis [83].

We make the following directionality assumptions for several bi-directional reactions:

- Fructose-biphosphate aldolase (FBA) that is part of mid-lower glycolysis is set towards catabolism [79].
- The bi-directional transports of magnesium and phosphate are both set towards uptake [80, 81].
- Acetate kinase (ACKr) and phospho-transacetylase (PTAr) are both set towards the acetate production, because acetate is one of the main by-products [61].
- The succinyl-CoA synthetase (SUCOAS) is set towards the production of succinate [61].

The polyphosphate kinases (PPK2r, and PPKr) are set towards the polyphosphate polymerization [82].

For some of intracellular metabolites, a corresponding transport reaction has not been biochemically characterized and does not appear in the *E. coli* iJO1366 and in our reduced model. However, these metabolites, unless they are highly polar or very large, are subject to passive diffusive transport through the cell membrane. Therefore, we explicitly added transport reactions for these metabolites that operate at least at basal level (10^-6 mmol/(gDW*h)).

3.4.2. Identification of alternative flux directionality profiles

As first step (Figure 3.1), in order to identify the reactions that are able to operate in both directions, flux variability analysis (FVA) was performed [16, 83]. If the system has a number z of bi-directional reactions, it could have up to 2^z FDPs. We enumerated the FDPs by adjusting the boundaries of the bi-directional reaction so that they can only operate in a unique direction. We define the coefficient of variability, CV_i, as:

$$CV_{flux,i} \approx \frac{UB_{flux,i} - LB_{flux,i}}{F_{flux,i}}$$

where, *UB* and *LB* are the upper and lower bounds respectively of the flux *i* derived using thermodynamic-based variability analysis (TVA) [16, 83]. *F* is the average of *UB* and *LB*. We define the flux variability score of each FDP as the Euclidean norm of the vector whose entries are the *CV* of each flux. The FDP with the highest flux variability score has the highest relative flexibility in terms of the allowable flux ranges.

3.4.3. Computation of reference and extreme steady states for alternative FDPs

For each of the identified FDPs, in step two (Figure 3.1), we sample the solution space of concentrations and fluxes without violating physiological, thermodynamic and directionality constraints. The convexity of these solution spaces enables us to efficiently generate sets of flux and concentration samples using the Artificial-Centering Hit-and-Run sampler in the COBRA Toolbox [8, 9, 109]. We perform principal component analysis (PCA) on the generated samples to select reference and extreme samples [91]. The first seven principal components (PC) were used as for the fluxes and the concentrations they covered above 90% of the sample variance. The reference sample is chosen so that its vector projections onto the seven PCs are minimal. We get the two extreme samples of a PC, PCmax and PCmin, by respectively finding the 0.1% top and the 0.1% bottom samples based on their magnitude of vector projections onto the given PC. Out of the 0.1% top and the 0.1% bottom samples we chose the samples that have the smallest magnitude of vector projections onto the other PCs (Appendix B3).

3.4.4. Analysis of alternative solutions between FDPs

3.4.4.1. Thermodynamic displacement analysis

Within each FDP, in step 3 (Figure 3.1), we compute the displacement of the reactions from thermodynamic equilibrium, Γ [52, 84, 87]. For a simple uni-uni reaction with a substrate *S* and a product *P*, the thermodynamic displacement, *Γ*, is defined as:

$$\Gamma = \frac{1}{k_{eq}} \frac{[P]}{[S]}$$

where, k_{eq} is thermodynamic equilibrium. More specifically, we first use the vector of the reference steady-state concentrations together with values of standard Gibbs free energies of reactions to compute *Γ*. For reactions with negative Gibbs free energy, $0 < \Gamma < 1$. For reactions that are far away from equilibrium *Γ* is close to 0, and for reactions near equilibrium $\Gamma \approx 1$. We then classify the reactions in terms of *Γ* in the following four classes: reactions that operate (i) near equilibrium (NE), $0.9 \leq \Gamma \leq 1$; (ii) near to middle equilibrium (NM), $0.5 \leq \Gamma \leq 0.9$; (iii) middle to far from equilibrium (MF), $0.1 \leq \Gamma \leq 0.5$; and (iv) far from equilibrium (FE), $0 \leq \Gamma \leq 0.1$. The information about *Γ* is important, as it is known that enzymes that operate near equilibrium do not have control over fluxes and concentrations in the network [84].

Within the MCA framework, Kaeser and Burns [88] define the concentration control coefficients, C_p^x, and the flux control coefficients, C_p^v, as the fracitonal change of metabolite concentrations and metabolic fluxes, respectively, in response to fractional change of system parameters. According to the log(linear) formalism [110, 111], we can derive C_p^x and C_p^v as:

$$C_p^x = -(NVE)^{-1}NV\Pi$$

$$C_p^v = EC_p^x + \Pi$$

where, *N* is the stoichiometric matrix, *V* is the diagonal matrix whose whole elements are the steady-state fluxes, *E* is the elasticity matrix with respect to metabolites and *Π* is the matrix of elasticities with respect to parameters. If we now consider a uni-uni reaction, *i*, with a substrate *S* and a product *P* we write its reaction rate v_i as follows:

$$v_i = V_{max} \frac{[1-\Gamma]\frac{[S]}{K_{M_S}}}{\frac{[S]}{K_{M_S}} + \frac{[P]}{K_{M_P}} + 1}$$

where, V_{max} is the maximum velocity at enzyme saturation, and, K_{M_S} and K_{M_P}, are the Michaelis constants of *S* and *P,* respectively. We define, as done previously by Hatzimanikatis and coworkers [50], the elasticities with respect to *S* and *P*, respectively, as:

$$\varepsilon_{v_i}^S = \frac{S}{v}\frac{dv}{dS} = \frac{1}{(1-\Gamma)} - \frac{\frac{[S]}{K_{M_S}}}{\frac{[S]}{K_{M_S}} + \frac{[P]}{K_{M_P}} + 1}$$

$$\varepsilon_{v_i}^P = \frac{P}{v}\frac{dv}{dP} = -\frac{\Gamma}{(1-\Gamma)} - \frac{\frac{[P]}{K_{M_P}}}{\frac{[S]}{K_{M_S}} + \frac{[P]}{K_{M_P}} + 1}$$

where, $\varepsilon_{v_i}^S$ and $\varepsilon_{v_i}^P$ are entries of the elasticity matrix *E*. Evidently, if the reaction is at thermodynamic equilibrium (i.e. *Γ* ≈ 1), the first terms of the elasticity terms $\varepsilon_{v_i}^S$ and $\varepsilon_{v_i}^P$ tend towards infinity and we consequently have no control with respect the considered enzyme. However, if the reaction is far away from thermodynamic equilibrium (i.e. *Γ* ≈ 0), the second terms of $\varepsilon_{v_i}^{S}$ and $\varepsilon_{v_i}^{P}$ can have impact on the elasticities, potentially resulting in control. Hence, it is essential to consider thermodynamic displacement with the kinetics in order to understand control at systems level. The elasticity matrix E is directly affected by Γ, and hence the control coefficients C_p^x and C_p^v will also be impacted.

3.4.4.2. Kinetic parameter sampling

We build populations of kinetic models for the computed vectors of the reference steady-state fluxes and concentrations. We integrate the information about the kinetic properties of enzymes available from the literature [92] and the databases [93, 94]. We use the reversible Hill kinetics [95] and convenience kinetics [96] for reactions with unknown kinetic mechanism. For kinetic mechanisms with no or partial information about their parameter values we sample the space of kinetic parameters by direct sampling of the degree of saturation of the active site of an enzyme considering one [84] or multiple enzymatic steps [52]. We then parameterize a population of kinetic models (Appendix B4 and B5), perform consistency verifications [51, 54, 84], and compute the flux and concentration control coefficients [84, 97]. The consistency verifications include a stability test of the model that verifies the Jacobian matrix has no eigenvalues with positive real part for the sampled set of parameters. This test relies on the assumption that the observed RSSs for flux and metabolite concentration are in a stable steady state at the observed time point. For more details about the ORACLE workflow for construction of large-scale kinetic models that are consistent both with thermodynamics and the observed data, the reader is referred to literature [48-52, 84-86, 98].

3.4.4.3. General statistics on FCCs across FDPs

We computed FCCs of the 275 enzymatic reactions with respect to their 275 enzymes as a quantitative output to compare how our MCA conclusions were consistent across the FDPs. Thus, we calculated the FCCs

$$C_{p_k}^{v_i} = \frac{\partial\, ln v_i}{\partial\, ln p_k} = \frac{p_k\, \partial v_i}{v_i\, \partial p_k}$$

where, v_i is the flux across a reaction *i* and, p_k is the concentration perturbation of an enzyme *k*. We then compute the mean of the FCCs, $\overline{C_p^v}$, across the kinetic models for an FDP (Table 3.1).

We considered the FCCs for fluxes that were larger than 0.01 mmol/gDW/h across all FDPs because we wanted to focus our study around the reactions that carry more significant amount of carbon (i.e., central carbon metabolism fluxes). Only 126 reactions satisfied this condition, which left us with 34’650 (126 reactions x 275 enzymes) FCCs (Appendix B3, Figure S7). To compare more significant FCCs, we only considered ones that had more than absolute 0.1 fold change across the 4 FDPs so that we focus on FCCs with significant control. This meant that we kept 1’263 out of the previous 34’650 FCCs (Appendix B3, Figure S8).

$\overline{C_{p_1}^{v_1}}$	$\overline{C_{p_2}^{v_1}}$	$\overline{C_{p_3}^{v_1}}$	$\cdots$	$\overline{C_{p_m}^{v_1}}$
$\overline{C_{p_1}^{v_2}}$	$\overline{C_{p_2}^{v_2}}$	$\overline{C_{p_3}^{v_2}}$	$\cdots$	$\overline{C_{p_m}^{v_2}}$
$\overline{C_{p_1}^{v_3}}$	$\overline{C_{p_2}^{v_3}}$	$\overline{C_{p_3}^{v_3}}$	$\cdots$	$\overline{C_{p_m}^{v_3}}$
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$
$\overline{C_{p_1}^{v_n}}$	$\overline{C_{p_2}^{v_n}}$	$\overline{C_{p_3}^{v_n}}$	$\cdots$	$\overline{C_{p_m}^{v_n}}$

Table 3. 1. Mean of flux control coefficients of a population of models.

3.4.5. Characterizing the distribution of kinetic parameters

3.4.5.1. Beta distributions

The kinetic parameter solution space is studied in step 4 (Figure 3.1) by sampling uniformly the degree of saturation of an enzyme’s active site as defined by Wang *et al.* [84]. We obtain distributions of scaled metabolite concentrations from this sampling and consequently, kinetic parameter distributions. The degree of saturation of an enzyme’s active site has a well-defined range from zero to one, allowing us to resort to parametric distributions for their characterization. Beta distributions provide an efficient way of quantitatively expressing variability over a fixed range by estimating its two parameters [112]. These parameters can be obtained and compared for populations of kinetic parameters generated with different operational configurations.

3.4.5.2. Implying prior beta distributions for sampling

In this work we compare how alternative steady states describing a physiology impact metabolic engineering conclusions. It is thus desirable to ensure that the sampled degrees of saturation of enzyme active sites are similar for the populations of kinetic models built around alternative steady states within FDPs in step 5 (Figure 3.1). Hence, we compute beta parameters describing the distributions of the kinetic parameters of a given RSS. These beta parameters are used to sample degrees of saturation of enzyme active sites for alternative steady states from similar density distributions using the prior samples. The Beta distribution parameters are implied within the ORACLE workflow as input for sampling degrees of saturations of enzymes when parameterizing new kinetic models. Beta distributions hence bias sample densities for the sampling of degrees of saturation states for an enzyme.

3.4.6. Analysis of alternative solutions within FDPs

We investigate in step 5 (Figure 3.1) how different flux profiles and metabolite concentration vectors, within FDPs, affect the populations of control coefficients. We separately studied the effects of the flux profiles and the metabolite concentration vectors, in order to decouple their effects on control coefficients. We take the reference steady-state concentration vector and we form the pairs with the extreme steady-state flux profiles computed in step 2 of the procedure (Figure 3.1). We then generate populations of kinetic models as described in step 3. In the generation of missing kinetic information, we use the distributions of kinetic parameters that have been characterized in step 4 for this FDP. This way, we obtain alternative populations of kinetic models that have in common the reference steady-state concentration and the distribution of kinetic parameters. We compare these populations of kinetic models together with the population of kinetic models that was computed in step 3 for the reference steady state of this FDP. This enables the assessment of the effects of alternative flux profiles within the FDP onto the control coefficients.

The effects of alternative values of concentrations on control coefficients are estimated in an analogous way, where we take the reference steady-state flux and we form the pairs with the extreme steady-state concentrations and we repeat the procedure discussed above. Taken together these two comparisons of alternative solutions allow us to identify sets of enzymes within a FDP whose control over the fluxes and concentrations in the network is robust both with respect to the alternative concentrations and fluxes. We also identify enzymes that are robust only with respect to the alternative concentrations or alternative fluxes.

3.4.7. Metabolic Engineering and Synthetic Biology Design

We next analyze in step 6 the results obtained in steps 3-5 (Figure 3.1) in the light of metabolic engineering and synthetic biology design. We single out the enzymes whose control over fluxes and concentrations of interest is consistent over all FDPs and within FDPs. In this step we can also design the experiments that would give sufficient information for discriminating alternative solutions between FDPs and within FDPs.

3.5. Appendix B

B1 Table. Thermodynamics-based metabolic flux analysis model. Spreadsheet giving the list of metabolites, reactions, variables, constraints and compartmentalization in the flux directionality profile 1.

B2 Table. Thermodynamic variability analysis. Spreadsheet giving the flux ranges and Gibb's free energy ranges for reactions and the metabolite log(M) concentration ranges.

B3 Supplemental Material. Additional information and supplemental figures.

B4 Text. System of ordinary differential equations describing FDP1. Non-linear kinetic model of *E.coli* for FDP1 giving the system of ordinary differential equations.

B5 Table. Reaction mechanisms describing the systems.

4. Statistical inference in ensemble modeling of cellular metabolism

4.1. Introduction

Kinetic models are becoming essential computational tools for studying the metabolism of organisms and understanding the dynamics of their cellular biochemical interactions [105]. However, the construction of kinetic models remains a challenging endeavor as large uncertainties are associated with the rate expressions describing all the reactions making up these cellular interactions [32]. Reaction mechanisms are rarely fully characterized for an organism, making it difficult to select appropriate rate expressions for reactions and, information on the parameter values required by these expressions is very scarce. Several ensemble modeling (EM) approaches that assign kinetic mechanisms to reactions, incorporate experimental data and sample unknown kinetic parameter values have emerged for generating populations of kinetic models [47, 65, 84, 113]. Yet, given the promising methodologies that exist for constructing populations of large-scale kinetic models, the community lacks in procedures for examining their uncertainty.

Kinetic models are generally constructed with a particular objective such as improving a substrate’s production, increasing cellular growth or advising experimentalists on physiological properties to measure [105]. Irrespective of the objective, comparing populations of variables - such as metabolic control analysis (MCA) sensitivity coefficients - computed from the kinetic models in order to derive conclusions is a fundamental step in computational modeling and engineering. To meaningfully compare populations of variables, it is important to consider their associated uncertainty, for which innumerous statistical approaches exist, making it sometimes a dubious task to select the “correct” method [114]. Despite originating from statistical mechanics, EM has only been employed in systems biology for two decades [55] and its use of statistical methods for managing uncertainty remains, to our knowledge, untapped.

Kinetic models of metabolism are generally constructed around a given steady state of interest, characterizing the system. Assuming that we know the metabolite concentrations, the flux values and the reaction mechanisms describing the system, we still have uncertainty in the kinetic parameter values. The Optimization and Risk Analysis of Complex Living Entities (ORACLE) framework handles this uncertainty by considering multiple alternative possible sets of models by sampling the parameter space until enough models are obtained, such that the mean and several other statistical modes of the model outputs converge [52, 84, 115]. Another EM approach generates populations of models to search for a unique model that best fits experimental data to construct a time-course dynamic model which describes the system [45, 47]. A workflow for constructing kinetic models considers populations and assesses statistical significance using univariate analysis of uncertainty in variables [65]. However, these frameworks for building kinetic models do not appear to consider multivariate statistical methods when accounting for uncertainty.

The purpose of this chapter is to suggest how statistical approaches can be used to consider uncertainty in kinetic models arising from sampling their kinetic parameter space by constructing simultaneous confidence intervals (CIs). To achieve this goal, as there is no unique approach for constructing simultaneous CIs, we review several methods. Each one comes with certain underlying assumptions and caveats that should be taken into consideration before application. We decided to compare how different approaches can be applied to our data and make recommendations on how such approaches can serve the community in attributing statistical significance to variables and handling uncertainty.

A simultaneous CI is a range that contains the true means of a set of variables with a fixed probability called coverage. Unlike well-known univariate CIs, simultaneous CIs account for the multiplicity of variables to achieve the coverage. Their constructions can be approximate or very technical, depending on the underlying distribution of the data. In this chapter, we expose three methods: Bonferroni's correction (BCI), the exact normal (ENCI), and the bootstrap (BootCI). We discuss their advantages, disadvantages, and assumptions, to suggest how these methods can be applied successfully for comparing variables.

For the application, we use a published kinetic model [116] of aerobically grown *E.coli* that was derived in chapter three. The ORACLE framework is used to compute populations of flux control coefficients (FCCs) derived with MCA. The FCCs represent the fold change in a specific flux with respect to the perturbation of an enzyme's activity of *p*=275 enzymatic reactions with respect to their enzymes. We studied the FCCs of *n*=50'000 kinetic models with the three previously mentioned methods for constructing simultaneous CIs and suggested a workflow (Figure 4.1) for applying them.

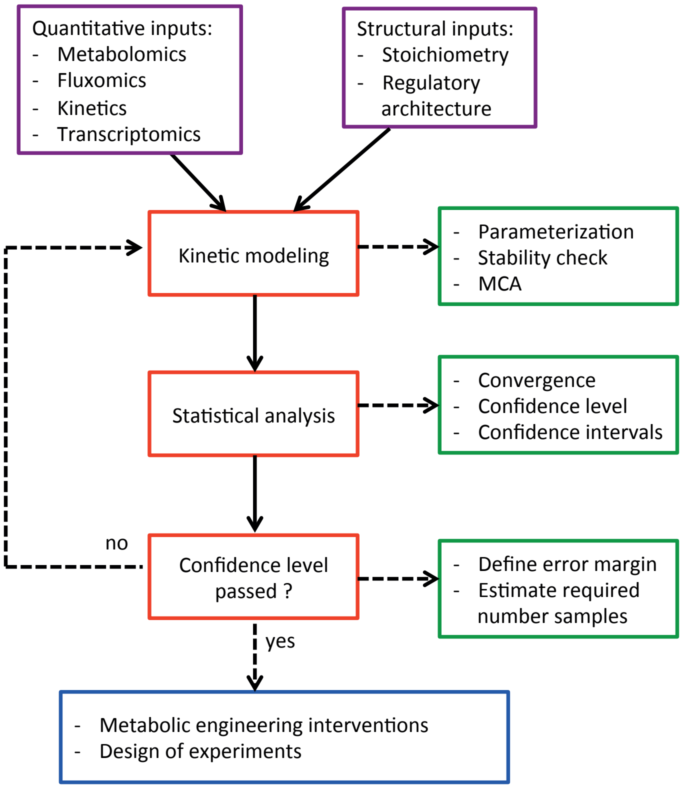


Figure 4. 1. Schematic diagram of workflow carried out in the study. Information about key steps of the workflow.

The rest of the chapter is organized as follows. In the Results and Discussion section we first demonstrate and discuss how three different statistical approaches can be used to construct simultaneous CIs and we then apply them to a case study (Figure 4.1). The Conclusion highlights our key findings. The Materials and Methods section provides further information about the statistical procedures used in this chapter. The supplementary material (Appendix C1) includes the algorithms used for constructing the simultaneous CIs.

4.2. Results and discussion

4.2.1. Kinetic model derivation

A reduced stoichiometric model of *E.coli* [116] was obtained with the redGEM and lumpGEM algorithms [57, 58] from the iJO1366 genome-scale metabolic model [77]. The reduced model is constituted of 277 enzymatic reactions and 160 metabolites distributed over cytosolic and extracellular space (Figure 4.2). Experimental flux and metabolite concentration data describing the optimal aerobic growth of *E.coli* were integrated within the thermodynamic formulation of the reduced stoichiometric model for aerobically grown *E.coli* from McCloskey *et al.* [61]. After integrating experimental data, alternative steady-state solutions could characterize the flux and the metabolite concentration of the studied physiology as the system was underdetermined (see Materials and Methods).

To construct kinetic models, we had to assume a steady state for the fluxes and for the metabolite concentrations [116]. As this study focused on comparing statistical methods for deriving CIs around the outputs of populations of kinetic models, we did not discuss the differences in the alternative steady-state solutions nor their biological implications. For a given case study, we assumed a solution for the fluxes and for the metabolite concentrations. We considered different case studies that were constructed around alternative steady states. For each case study that we considered, we had to sample the kinetic parameters for the kinetic mechanisms describing the molecular interactions of the cell.

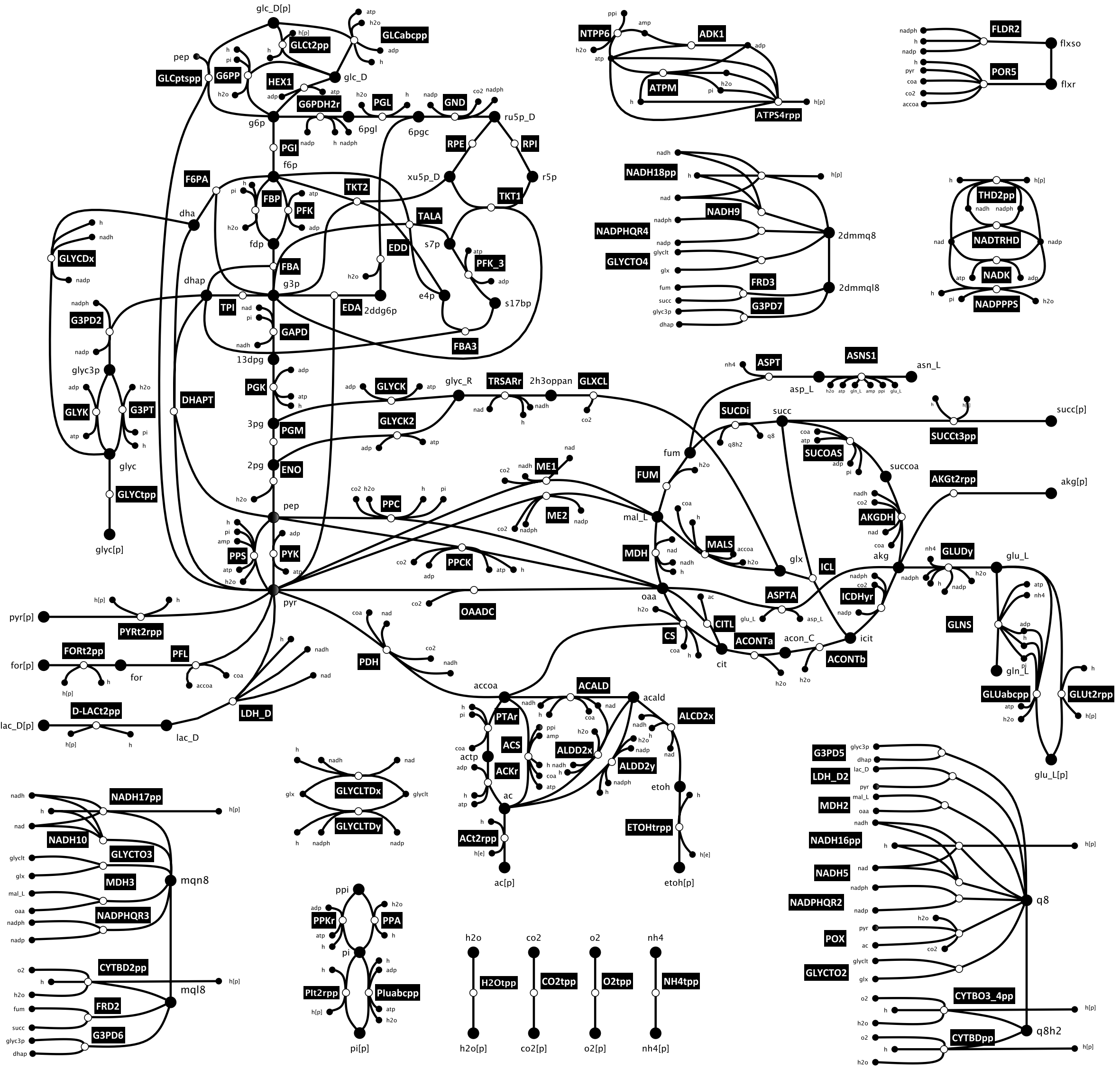


Figure 4. 2. ***E.coli*** **network diagram illustrating the core topology studied. Diagram does not include all the reactions of the systems. Full reaction and metabolite names are given in the supplementary (Appendix B1).**

We assigned reversible Michaelis-Menten enzyme kinetics (Uni-Uni, Uni-Bi, Bi-Bi, Bi-Ter etc…) to the reactions of the metabolic network [92]. If, for some reactions, the mechanism was not known, we resorted to using generalized reversible Hill kinetics [95] or convenience kinetics [96]. This resulted in a kinetic space of 1411 enzyme saturations for which we had to sample 1411 corresponding K_m values. We sampled uniformly all the enzyme saturations between 0 (non-saturation) and 1 (full saturation) using the ORACLE framework [84]. The local stability of models around the given steady state was tested and only stable models were kept (see Materials and Methods). As an inherent part of the ORACLE workflow, we compute control coefficients for each stable model with MCA.

4.2.2. Uncertainty in flux control coefficients

FCCs derived from MCA have been used for metabolic engineering purposes to give insight on the rate limiting steps of a metabolic network. Hence, it has been desirable to compare FCCs to find which enzyme could be edited and to achieve certain target metabolic state of the cell. For a given flux profile and metabolite concentration vector, referred to as case 1, we considered n=50’000 stable kinetic models generated with the ORACLE workflow. Alternative cases are used in the next section but, for the illustration of the different simultaneous CI in this section, we only focused on case 1. We consider p=275 FCCs of glucose uptake (GLCptspp) to determine which enzymes have the most control on it and could be of interest for editing. GLCptspp was chosen as an example, rather than for any specific biological reasons/objective. The p=275 FCCs will be considered as the variables, for which we have n=50’000 observations. To get some insight into which variables have the largest population mean μ, Figure 4.3 shows sample means sorted by absolute value.

4.2.3. Confidence intervals

We considered the four methodologies for building CIs presented in Materials and Methods; one without correction, and three that account for the simultaneous coverage level. We used the case 1 as an example to study and compare the CIs. We pre-processed case 1 data by removing variables that had a standard deviation below a tolerance level of 10^{-9}.

When we applied the Bonferroni’s corrections to case 1, we noted that the CI ranges were considerably larger than the ones obtained via t-distributions without correction (Figure 4.3A and 4.3B). This was expected as the coverage levels were adjusted for simultaneity and thus more conservative.

Regarding the BootCIs, they were generally slightly smaller than the ones derived with the exact normal method (see Figure 4.3C and 4.3D). This was expected, as the BootCIs are less conservative than the exact normal ones. Nevertheless, when the distribution was heavily skewed, the asymmetric CI could be considerably larger on one side of the data point. The happened with the oxygen transport (O2tex), phosphofructokinase (PFK), phosphate transport (PItex) and carbon dioxide transport (CO2tex) (Figure 4.4). As the bootstrapping method uses the observed data to derive the CIs, they appear more representative and adapted to the studied data.

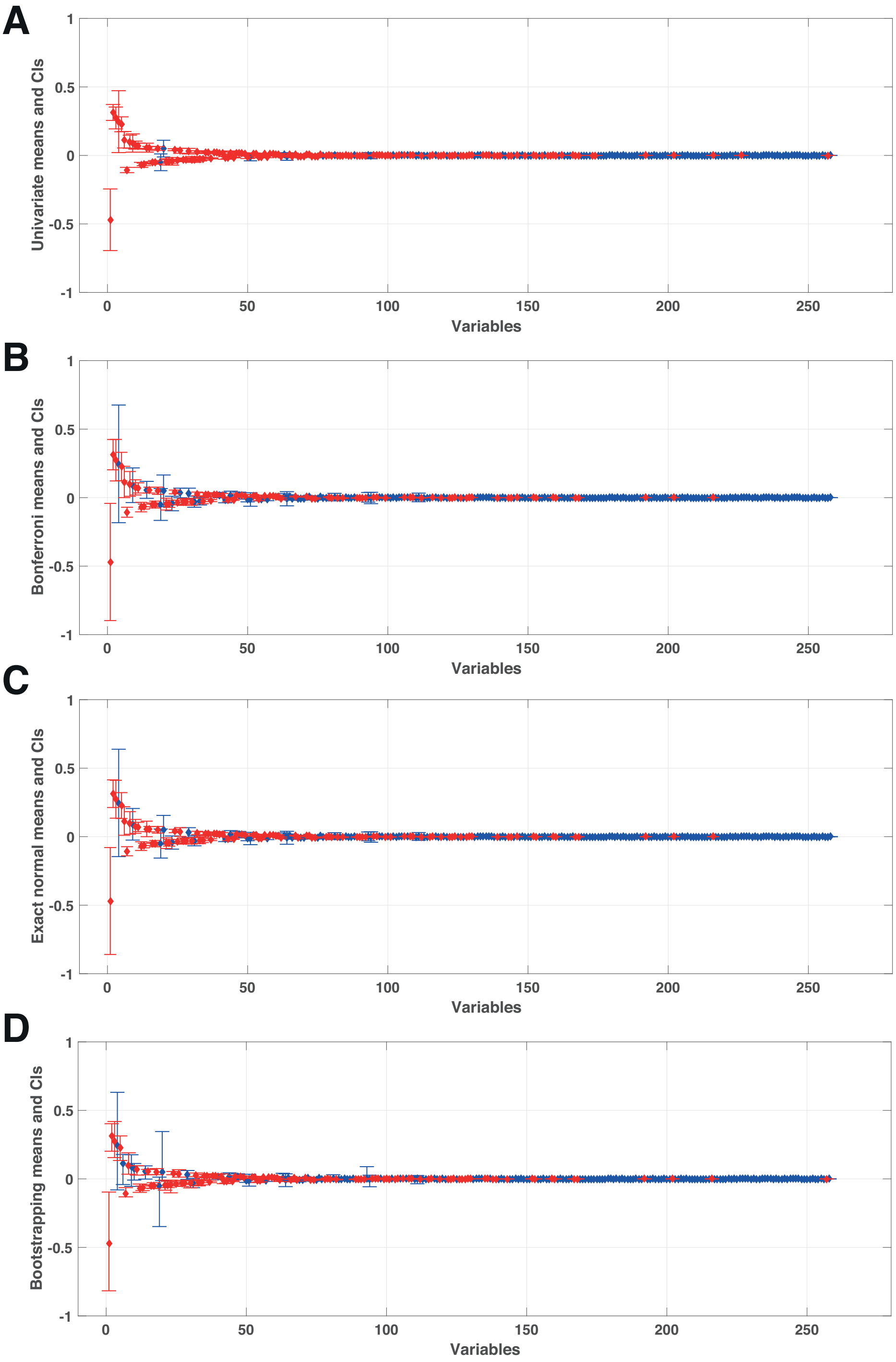


Figure 4. 3. Top control coefficients for glucose transport (GLCptspp). The diamonds indicate the mean of the FCCs in decreasing order of absolute mean. CIs were derived using (A) univariate t-test, (B) Bonferroni, (C) exact normal and (D) bootstrapping (see Materials

and Methods). The lower and upper whiskers correspond to the CIs. The CI is blue if it contains zero and, red if it does not.

The computational times for obtaining BCIs, ENCIs and BootCIs were recorded as 0.12 s, 0.47 s and 3520 s, respectively (Mac Pro, 2.7 GHz 12-Core Intel Xeon E5, 64 GB 1866 MHz DDR3 ECC). As expected, the computation of BootCIs was considerably longer than both the Bonferroni's and the exact normal methods. This is due to its intense re-sampling as exposed in Materials and Methods.

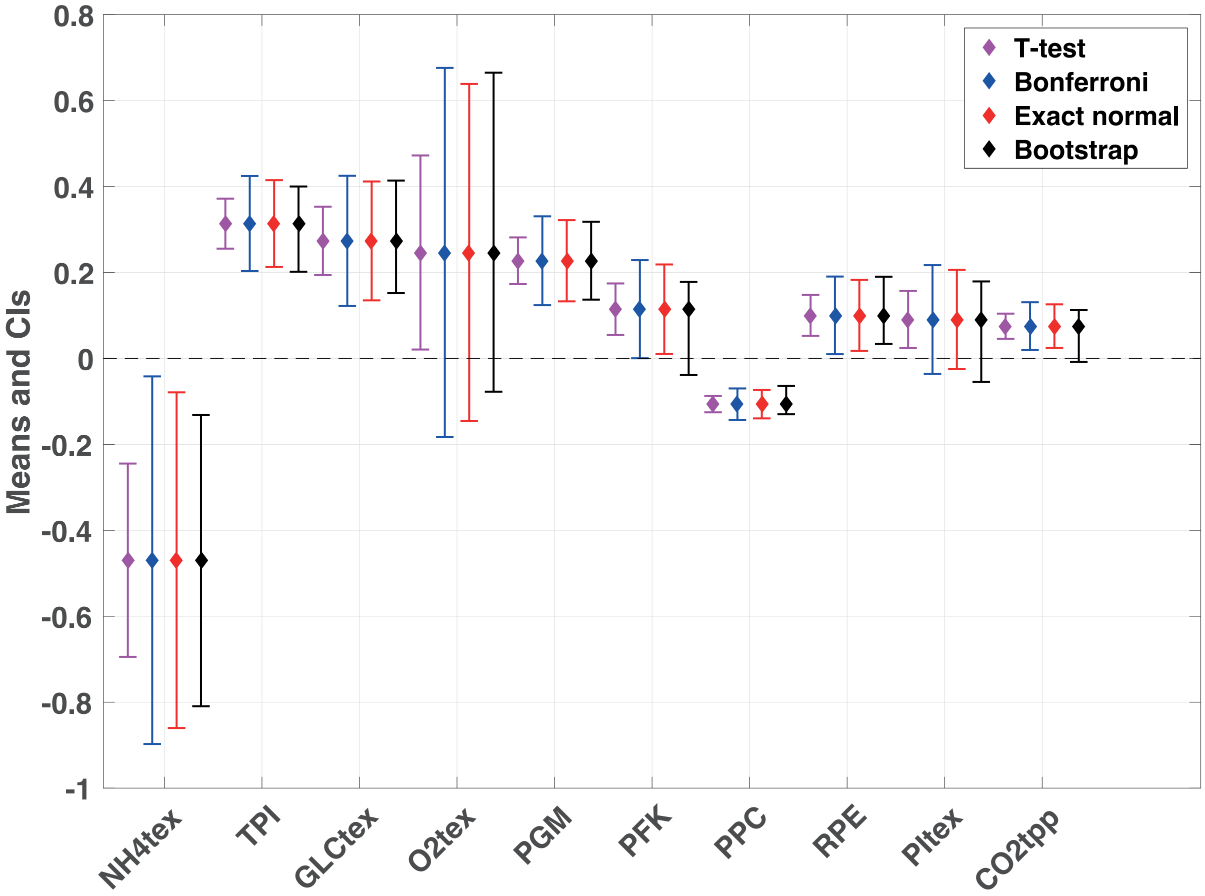


Figure 4. 4. Top flux control coefficients of glucose uptake (GLCptspp) with confidence intervals determined by different statistical approaches. The top 10 FCCs based on absolute mean are reported with diamonds. The whiskers indicate the CIs for univariate t-test (magenta), Bonferroni (blue), exact normal (red) and bootstrapping (black). The reader is referred to the Materials and Methods for technical details on CI computation.

Due to the underdetermined nature of our system, alternative steady-state solutions could describe the experimentally observed *E.coli* physiology. Consequently, we constructed populations of kinetic models around alternative solutions and, we then wanted to compare their MCA outputs. Studying these outputs can help elucidate why steady-state solutions affect the MCA-based control patterns and metabolic engineering decisions. Fortunately, these three statistical techniques presented in this chapter can be applied for building simultaneous CIs for the average difference between two data sets for our case study (see also Appendix C1).

4.2.4. Case study: mean difference confidence intervals

We were interested in studying four different steady-state solutions that can characterize the physiology of aerobically grown *E.coli*. We constructed populations of 50'000 stable kinetic models for these four cases using the ORACLE workflow. The first case referred to as case 1 was already presented previously to demonstrate the different methods for constructing CIs. The three other cases will be referred to as case 2, case 3 and case 4, as we do not discuss their biological differences in this chapter. These four cases correspond to different flux directionality profiles (Figure 3.2) that were discussed in the previous chapter in detail. Here, we want to compare how different the FCCs for GLCptspp were for these four cases.

In order to make this case study more comprehensible, we made a prior selection of the FCCs for GLCptspp that we wanted to compare. For each case, we built the bootstrapped simultaneous CIs and kept the seven FCCs with the largest absolute value in mean amongst those significantly different from zero. The union of these top seven FCCs of the cases was selected for the comparisons, resulting in 15 FCCs to be compared. The bootstrap was selected because it appeared the most appropriate technique to study data that can be highly skewed. Both the Bonferroni's and exact normal methods could have been used also. Since we wanted to compare these 15 FCCs between all the cases, this resulted in 90 comparisons overall (15 x 3 x 2).

The three statistical methods exposed in Materials and Methods were used to construct CIs for these 90 comparisons. Again, the bootstrapping approach is expected to be the most appropriate because of the aforementioned skewedness of the data. Overall, as shown on Figure 4.5, 45 comparisons were significant based on the bootstrapping approach. In comparison, The Bonferroni's and the exact normal methods resulted in 44 and 39 significant comparisons, respectively (Appendix C2 and C3). Since the variables have little correlation in this case study, the complexity of the exact normal over the BCIs seems not to be needed.

We also noted that the comparison of nicotinamide adenine dinucleotide kinase (NADK) between cases 1 and 2 and between cases 1 and 3 appeared to be significant when using the bootstrapping approach, whereas the other two approaches would suggest it is insignificant. This is an excellent example evidencing that the approaches relying on the normality assumption may lead to different conclusions than the bootstrapping method, when having to deal with skewed distributions. Yet, overall, there were no major differences in the widths of the CIs derived for the 90 comparisons for these three statistical methods.

Regarding the overall results in the applications, we observed very few differences between statistical methods, which gave us a preference for the Bonferroni's method due to its simplicity. The absence of any major differences between these methods can be explained by the fact that the correction for simultaneity is the first and most important aspect, before accounting for the dependence and for the skewness of the distribution. This was probably due to the very large number of variables/comparisons in our considered examples. In addition, it was evident that the main factor driving the CIs was the standard deviation of the distributions. We had a clear example of variance inhomogeneity between the variables. If all these techniques take reasonably well into account this inhomogeneity, it is not surprising to see that most variation from one CI to another is indeed due to the standard deviation. This was probably why taking one technique or another did not change the practical results too noticeably.

It should also be mentioned that the Bonferroni's method, in its full simplicity, allows a sample size calculation to estimate the number of samples required to achieve a certain level of confidence (see Materials and Methods, Section 5). This is an a posteriori calculation that is done based on the samples that we already have. For instance, in order to attain BCIs that have a maximal margin of error of 0.1, we would require around 1.9 million samples for our case studies based on these 90 comparisons that we performed here. Obviously, this estimated number of samples required is subject to the basic and conservative assumptions of the Bonferroni's method and only serves as an indication.

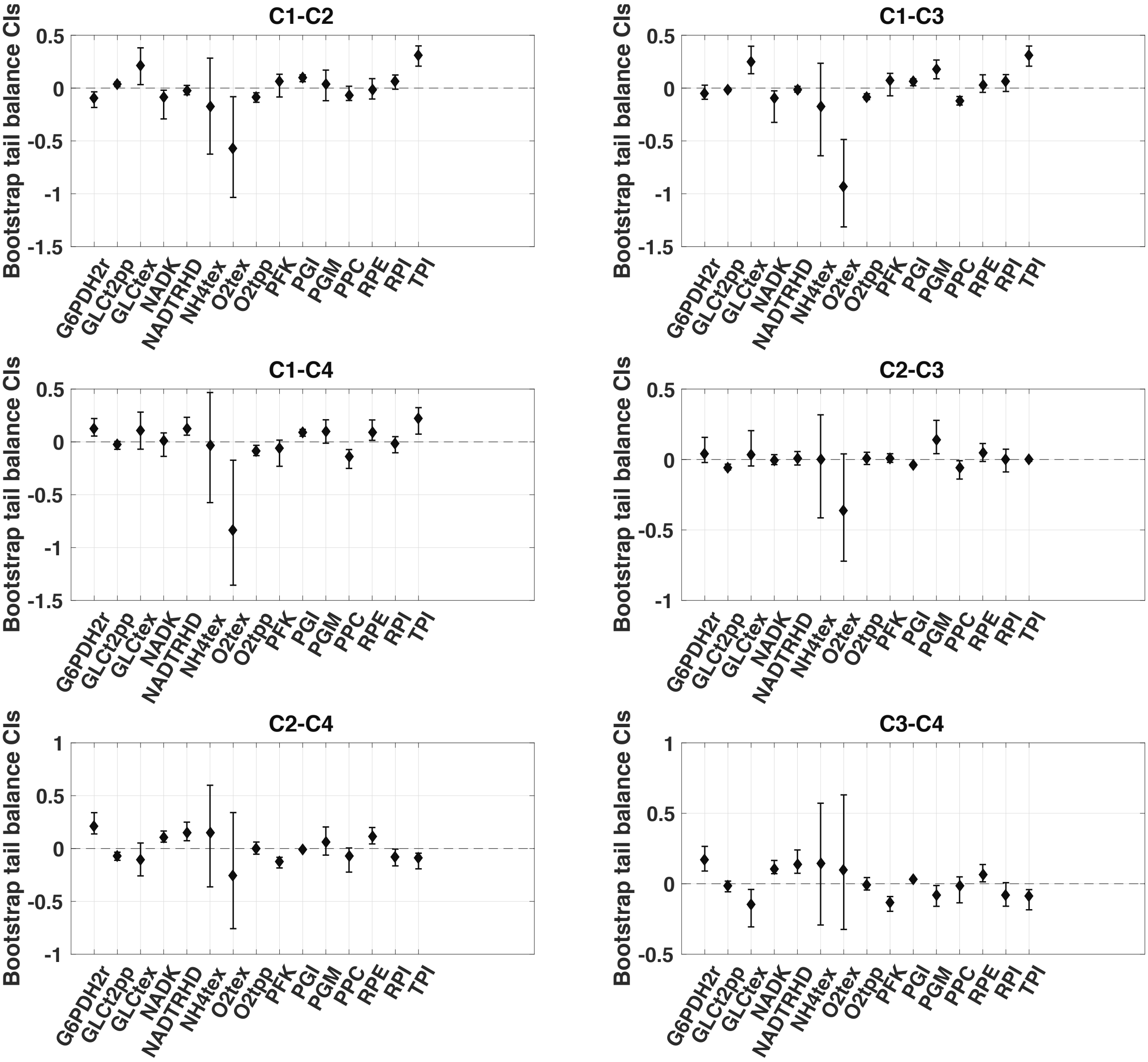


Figure 4. 5. Case study: differences of means using bootstrapping. Comparison of the differences in means of 15 FCCs for GLCptspp across 4 cases using the bootstrapping method (see Materials and Methods). The whiskers indicate the CIs and the diamonds report the estimates of the differences in means. The tests were carried out globally on the 90 estimates even though we report each case comparison as a separate plot.

However, we noted that bootstrapping provided certain minor advantages over the other methods, at the cost of higher computational efforts, particularly when the distributions are heavily tailed or asymmetric. Hence, if the additional computational costs are not too significant, for security, it may be more beneficial to apply the bootstrapping approaches when dealing with these kinds of data sets. As this was clearly the case for us, and we had the computational resources, it was certainly worth investigating and applying the bootstrapping approach for our results. Should the bootstrapping approach be too complex to implement computationally, it is worthwhile considering the exact normal method over the Bonferroni's one in the presence of high dependencies between the variables.

4.3. Conclusion

We hereby introduced, to our knowledge, the first computational workflow for assigning simultaneous CIs to populations of FCCs derived from kinetic models of metabolism. This work studied how alternative statistical approaches can be applied for computing CIs for the MCA outputs of populations of non-linear kinetic models of the metabolism of aerobically grown *E.coli*. We investigated the differences in three distinct methods – Bonferroni's correction, exact normal, and bootstrap – in calculating simultaneous CIs and discussed their particularities and assumptions. We evidenced with FCCs that we could successfully use these three methods to build CIs for populations of models. There were no considerable differences in the CIs derived using the three methods in the exposed data. However, the Bonferroni's correction was remarkable for its simplicity and for its readiness for estimating sample sizes required for achieving certain confidence level. We highlighted that the bootstrapping approach, although more complicated computationally and algorithmically, provided certain clear advantages when handling data with highly asymmetric and/or skewed distributions. Independent of the method used, it was crucial to consider the correction of CIs for their simultaneity. Hence, we propose a workflow (Figure 4.1) that can be used to construct CIs for the outputs – not only limited to control coefficients derived from MCA – of populations of kinetic models.

The statistical methods developed in this chapter allowed us to quantify uncertainty in control coefficients by using three different methods for deriving simultaneous confidence intervals. However, it is of interest for the community to reduce this uncertainty in control coefficients. In the next chapter, we study how sensitivity analysis can be used to source kinetic parameters that contribute the most to the uncertainty associated with selected control coefficients.

4.4. Materials and methods

4.4.1. Model reduction

The stoichiometric model for this study was obtained from the *E. coli* iJO1366 [77] using redGEM and lumpGEM, a set of frameworks for developing core models consistent with their genome-scale counterparts [57, 58]. In the process of reducing the *E. coli* iJO1366 [77], we defined the carbon source, the content of the cell media and the metabolic sub-systems of interest for the study, allowing us to derive context-specific models. Glucose was the sole carbon source and the subsystems of interest were ones from the central carbon (glycolysis/gluconeogenesis, pentose phosphate pathway, citric acid cycle, pyruvate metabolism and glyoxylate metabolism). We integrated the omics data available for the physiology of optimally grown *E. coli* under aerobic conditions from McClosekey *et al.* [61] into the mixed integer linear programming formulation of the model thermodynamics [83].

4.4.2. Kinetic parameter sampling

Ensembles of kinetic models were constructed around the computed vectors of the reference steady-state fluxes and concentrations for each case [116]. We integrated the applicable information about the kinetic properties of enzymes available from the literature [92]. The reversible Hill kinetics [95] and the convenience kinetics [96] were used for reactions with unknown kinetic mechanism. When no or partial information is available about kinetic parameters, we sampled the space of kinetic parameters by direct sampling of the degree of saturation of the active site of an enzyme considering one [84] or multiple enzymatic steps [52]. Ensembles of kinetic models were parameterized in order to perform consistency verifications [51, 54, 84], and compute MCA control coefficient [84, 97]. As part of consistency verifications, the ensembles of models were tested for not having any eigenvalues with positive real part. We assumed for this test that the observed reference steady-state solutions for flux and metabolite concentration were in a stable steady state at the observed time point. Further details about the ORACLE workflow for construction of large-scale kinetic models that are consistent both with thermodynamics and the observed data are available in literature [48, 51, 52, 84-86, 98, 115].

4.4.3. Simultaneous CIs for variable significance

A CI is an interval that contains the population mean μ with a probability of $1 - \alpha$, called the coverage. The population mean can be thought of the limit sample mean as *n* tends to infinity. The CI is built from the sampled data and is thus random. The coverage is to be understood as the proportion of times the CI would contain μ if the sampling were repeated a large number of time.

The variable significance is judged by its population mean μ estimated by the sample average. This estimate is tainted by uncertainty due to the variation of the data. This uncertainty is quantified by CIs. Because of the equivalence between statistical test and CI, to be of real importance, a variable sample average should be large in absolute with a CI bounded away from zero to ensure that this large estimated value is due to pure chance. In the following, several constructions of CI are presented.

4.4.3.1. Univariate and simultaneous CI

CIs can be built using innumerous techniques and for any parameters. The most well known CIs for the mean are univariate and based on the t-distribution. To account for the variability, univariate CI at level of $1 - \alpha$, for α = 5%, are added around each sample mean (see Figure 4.3). Checking that the CI contains 0 is equivalent to making a statistical test that $\mu = 0$ at level α. All technical details are recall in supplementary materials (Appendix C1).

Used as such, univariate CI are misleading since a correction for the fact that we inspect *p* variables is needed. This need, well-known for multiple testing [117], is the same for CIs. Indeed, the simultaneous coverage of several CIs, which is the probability of containing all population means, may be much lower than each univariate coverage. In the remaining of this section, we present three ways to build corrected CI, called simultaneous CIs.

4.4.3.2. Bonferroni's simultaneous confidence interval (BCI)

The Bonferroni's correction, probably the most used method, guaranties the simultaneous coverage $1-\alpha_S$ by dividing the univariate α levels by *p*, giving $\alpha = \alpha_S / p$. For example, with variables, each CI is built at $\alpha = .025$ and the simultaneous coverage is $(1-\alpha)^2 = .975^2 = .951$. On a larger scale, for example with our 275 variables, without correction the simultaneous coverage would be $.95^{275} \approx .00$, if the variables are all independent. Hence, it is almost sure that at least one of the population means is not contained in the corresponding CI.

This correction is approximate and correct only if the variables are all independent (see Appendix C1). Otherwise, it is often too conservative which means that the CIs are too wide [118].

An important aspect is that, even when the Bonferroni's correction (or any other) is appropriate, the univariate coverage will be $1-\alpha > 1-\alpha_S$. Thus, taken individually, each CI is conservative. This is the cost of having simultaneous correction. This is illustrate in Figure 4.3B.

The independence assumption of the Bonferroni's correction is not often satisfied, as shown in our case (Appendix C4). Hence, we felt encouraged to consider alternative approaches that account for the dependency of the variables being compared.

4.4.3.3. The exact normal (ENCI)

The exact normal method [119] attempts to release the Bonferroni's assumption of independence between the CIs. The method uses multivariate normal distributions $N_p(0, \Gamma)$ to correct for the dependencies of the p variables using an estimate of Γ, the correlation matrix of the observations. If the variables exhibit dependence, the resulting simultaneous CIs are expected to be smaller than the ones derived with Bonferroni's correction. The price to pay is in terms of computation and mathematical complexity. For the technical details, see supplementary materials (Appendix C1).

Both the exact normal and the Bonferroni's correction rely on the normal distributions assumption for constructing CIs. However, extreme observations and asymmetry in the data justify using methods relaxing this assumption.

4.4.3.4. Bootstrapped simultaneous CI (BootCI)

Originated from Beran's work [120], the BootCIs generalize the exact normal by relaxing the normality assumption. The approach is based on the re-sampling of the data in order to estimate a root statistic distribution. Coupled with a pre-pivoting technique and tail balancing, and under some technical assumptions, the method provides asymmetric simultaneous CIs with

- The correct target simultaneous coverage
- Equal marginal coverages
- Outside tail balance, i.e. the same probability on both sides out of the CIs.

The bootstrap assumption is lighter than the normality assumption but it has to be valid. Unfortunately, it cannot be validated in practice and remains an a priori assumption.

The BootCIs are the most general available without any further assumption to our knowledge. They are always more correct than Bonferroni’s and exact normal in that if the assumptions of the former are valid, the bootstrapped one are also valid. The price to pay is yet another level of technicity and computation complexity. For the technical details, see supplementary materials (Appendix C1).

4.4.4. Confidence intervals for comparison of cases

The comparison of two cases is made by building CIs on the difference of their means. When cases are compared along several variables, simultaneous CIs must be used and can be built using the three methods seen in Section 0. In applications, simultaneous CIs are used for multiple comparisons (for example see [121] for a detailed treatment). Because of the correction for simultaneity, the variables along which the cases differ can be tested: the differences are significant whenever zero does not belong to the interval (see Figure 4.3 for the application). The mathematical details are reported in supplementary materials (Appendix C1).

4.4.5. Sample size calculation

The use of confidence intervals and of power analysis are well known in the computation of the required sample size (for example see [122] for a good overview in clinical research context). The general concept is that the length of a CI diminishes when the sample size increases. Since this length is measuring the uncertainty on the corresponding mean, the required sample size to achieve a given length can be computed.

However, the sample size computation requires prior knowledge or prior data gathering. Indeed, the length of the CI depends also on the standard deviation that has to be guessed or estimated beforehand. In our application, we thus make an a posteriori sample size computation based on the estimated standard deviation from the available sample. We also used the BCI because it is the only method allowing the formulation of an explicit sample size calculation (Appendix C1). Even if approximate, this calculation would be quite demanding with the other methods, if not intractable for the bootstrapping approach.

4.5. Appendix C

C1 Supporting Information. Algorithms and further details for deriving confidence intervals.

C2 Figure. Case study: differences of means using bonferroni method. Comparison of the differences in means of 15 FCCs for GLCptspp across 4 cases using the bonferroni method (see Materials and Methods). The whiskers indicate the CIs and the diamonds report the estimates of the differences in means. The tests were carried out globally on the 90 estimates even though we report each case comparison as a separate plot.

C3 Figure. Case study: differences of means using exact normal method. Comparison of the differences in means of 15 FCCs for GLCptspp across 4 cases using the exact normal method (see Materials and Methods). The whiskers indicate the CIs and the diamonds report the estimates of the differences in means. The tests were carried out globally on the 90 estimates even though we report each case comparison as a separate plot.

C4 Figure. Heatmap of correlation matrix for case 1 variables.

5. Global sensitivity analysis of control coefficients derived with metabolic control analysis

5.1. Introduction

New computational frameworks are enabling the construction of genome-scale kinetic models that are consistent with stoichiometric, thermodynamic and physiological constraints [123]. Despite the advances of experimental methods for estimating kinetic parameters, significant uncertainty in their nominal values and ranges remains. Adding to this uncertainty, the number of kinetic parameters that characterize the system increases with the size of kinetic models. To overcome the problem of assigning unique kinetic parameter values, one solution is to sample the kinetic parameter space, generating multiple alternative models [26, 45, 47, 124]. This uncertainty in kinetic parameters, as demonstrated by Andreozzi *et al.*, can result in kinetic models with contradicting properties and conclusions [49]. Their work mirrors the findings by Gutenkunst and coworkers who suggest that models in systems biology are 'sloppy models' and that there are usually only few parameters that affect model outputs [66]. Hence, information about which kinetic parameters, if measured experimentally, would reduce the most the uncertainty in kinetic model outputs is essential for the design of experiments and improving model predictions.

Different sensitivity analysis approaches exist for the identification of model inputs that contribute the most to the uncertainty of model outputs [67-69, 125]. Variance-based global sensitivity analysis (GSA) approaches are the most established techniques for performing sensitivity analysis on systems described by nonlinear ordinary differential equations [67]. Their ability to estimate variance-based sensitivity indices allows them to quantify parameter-parameter interactions and their impact on model outputs. The model itself is treated as a "black box", making these methods applicable to any type of model.

Kiparissides and Hatzimanikatis developed a GSA procedure for analyzing genome-scale stoichiometric models that have thermodynamic constraints [126]. However, it appears to our knowledge that GSA approaches have not been applied to large-scale nonlinear kinetic models.

Both local and global sensitivity analysis methods have been applied to kinetic models of smaller scale, extending to reaction networks the size of pathways and subsystems [67, 127]. Performing GSA on genome-scale kinetic models of metabolism is challenging and can be computationally very expensive. We utilize the Optimization and Risk Analysis of Complex Living Entities (ORACLE) framework to construct population of kinetic models [26, 124]. ORACLE efficiently samples the kinetic parameter space using enzyme saturations (Methods) and computes Metabolic Control Analysis (MCA) sensitivity coefficients. The MCA outputs are mathematically derived from the sampled enzyme saturation levels that can be considered as inputs to the system. Hence, we developed a variance-based GSA approach for assessing the sensitivity of MCA outputs to the input enzyme saturation levels.

We use an *E.coli* model that was reduced [57] from the iJO1366 genome-scale model [77] to perform the sensitivity analysis. The model is constituted of 271 enzymatic reactions, 247 lumped reactions and 160 metabolites, resulting in a total of 3083 enzyme saturation levels to be sampled. It would be very computationally intensive to perform GSA with respect to all the enzyme saturation levels due to the size of the model. Variance-based sensitivity indices to study the total and first order effects of input parameters were computed based on Sobol indices [128]. Higher order effects would require considerably larger computational efforts. Hence, we first developed a workflow for identifying parts of the network that contribute the most to the variance of model outputs using a coarse-grain sampling approach. Once we have identified these parts of the network, we can perform a fine-grain sampling of the input parameters to identify the ones contributing the most to the variance of model outputs. The workflow was used to rank input parameters based on their contribution to the variance of MCA outputs and can be applied to different large-scale nonlinear models.

5.2. Results and Discussions

We have been developing a workflow (Figure 5.1) that performs variance-based global sensitivity analysis on metabolic control analysis (GMCA) of large-scale kinetic models and some results are presented here. The workflow required as an input a kinetic model for a given physiology and certain study specifications. These specifications, also referred to as the study scope, described which parts of the network we want to study and the resolution of the sensitivity analysis. By resolution we refer to whether we are considering parameters in groups or several parameters independently. The first step of the procedure involved sampling the kinetic parameters so that we can characterize the solution space. After this, subject to the study specifications, new populations of kinetic parameters were sampled. Then, these populations of kinetic models were used to compute Sobol indices (Methods) for desired model outputs. Based on these sensitivity indices, we determined parameters that were responsible for the variance of the kinetic model outputs, which in this case were the flux control coefficients (FCCs) computed with MCA. We then fixed some of these kinetic parameters in order to validate the findings. These findings could also be used to devise further experimental designs and/or metabolic engineering decisions.

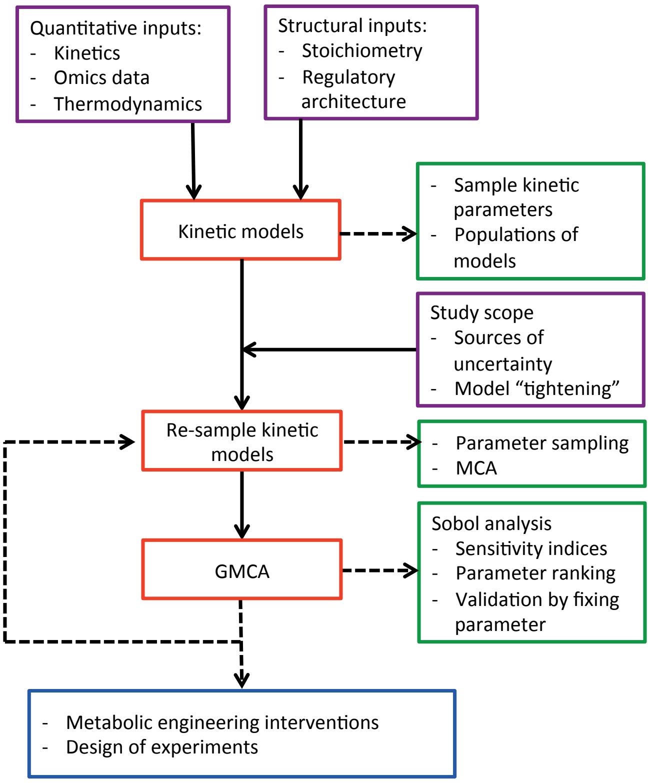


Figure 5. 1. Global sensitivity analysis workflow for characterizing sources of variability in large-scale kinetic models. Diagram providing details of the various steps required for the characterization of parameters responsible for variance in kinetic models and their outputs.

5.2.1. Kinetic model

We used a kinetic model, referred to as "D1" in chapter two, describing the physiology of aerobically grown *E.coli* for the purpose of this study (Figure 2.1). Although being the smallest model presented within chapter two, it still had 3083 Km values to be sampled. To characterize the kinetic parameters in our system, we used the ORACLE workflow for constructing populations of kinetic models. The same steady states for the metabolite concentrations and the metabolic fluxes as well as the kinetic mechanisms were used for this kinetic model as the one described in chapter two and its supplementary material (Appendix A2 and A6).

5.2.2. Uncertainty from the pentose phosphate pathway

We decided to focus on the pentose phosphate pathway (PPP) for this analysis because it contained 12 reactions that overall entail 38 Km parameters, making the scope of the study more tractable (Figure 5.2). We considered the parameters for these reactions in groups in order to find which reactions contributed the most to the variability of the kinetic models (Table 5.1). Since these reactions involved different numbers of parameters, it could be fair to expect - even without accounting for their location within the network topology - that they have different impact on the variability of the kinetic model outputs. We performed GMCA on the populations of kinetic models in order to rank the PPP reactions based on their contributions to uncertainty in FCCs (Methods).

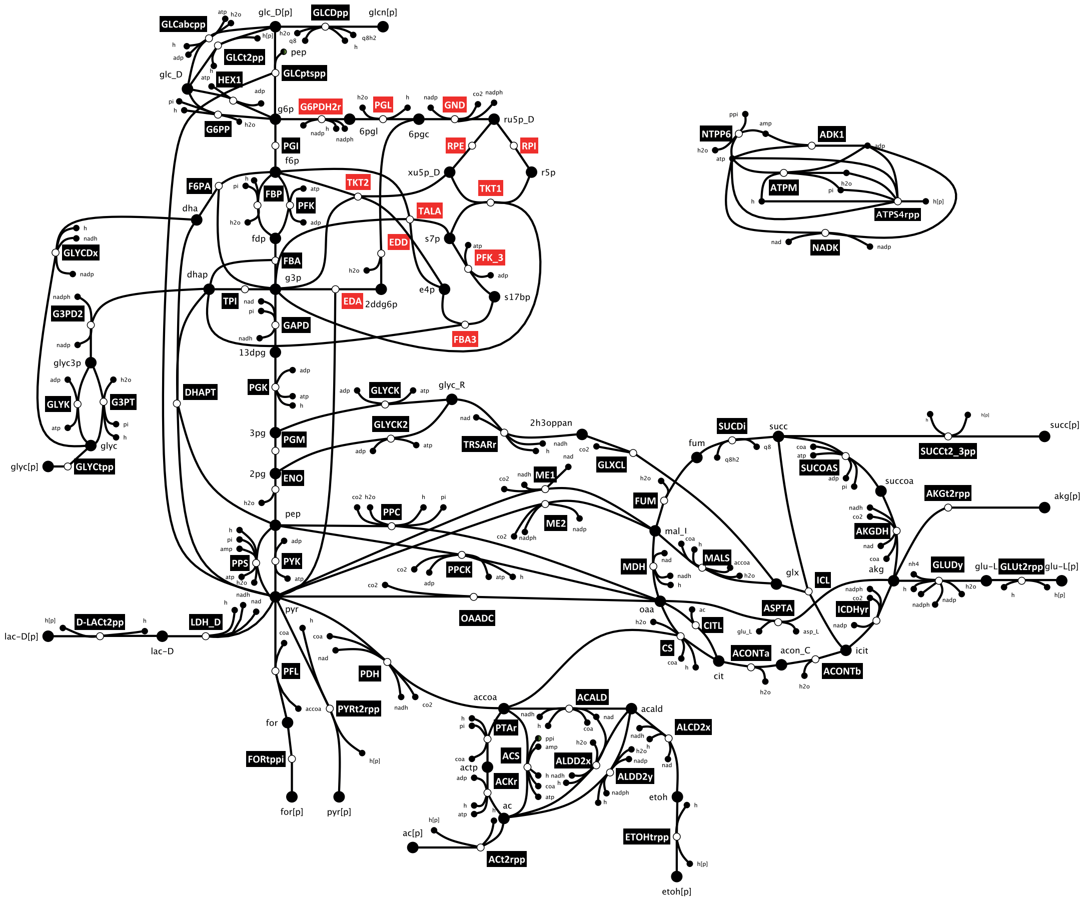


Figure 5. 2. *E.coli* network diagram illustrating the reactions of the kinetic model. The reactions indicated in red correspond to the PPP reactions whose uncertainty we studied using variance-based global sensitivity analysis. Diagram does not include all the reactions of the system.

Reaction	Number of Km parameters	Km Parameters
EDA	3	$K_{\mathrm{EDA,2ddg6p}}$, $K_{\mathrm{EDA,g3p}}$, $K_{\mathrm{EDA,pyr}}$
EDD	2	$K_{\mathrm{EDD,6pgc}}$, $K_{\mathrm{EDD,2ddg6p}}$
FBA3	3	$K_{\mathrm{FBA3,s17bp}}$, $K_{\mathrm{FBA3,dhap}}$, $K_{\mathrm{FBA3,e4p}}$
G6PDH2r	4	$K_{\mathrm{G6PDH2r,g6p}}$, $K_{\mathrm{G6PDH2r,nadp}}$, $K_{\mathrm{G6PDH2r,6pgl}}$, $K_{\mathrm{G6PDH2r,nadph}}$
GND	4	$K_{\mathrm{GND,nadp}}$, $K_{\mathrm{GND,6pgc}}$, $K_{\mathrm{GND,nadph}}$, $K_{\mathrm{GND,ru5p\text{-}D}}$
PFK_3	4	$K_{\mathrm{PFK\_3,atp}}$, $K_{\mathrm{PFK\_3,s7p}}$, $K_{\mathrm{PFK\_3,adp}}$, $K_{\mathrm{PFK\_3,s17bp}}$
PGL	2	$K_{\mathrm{PGL,6pgl}}$, $K_{\mathrm{PGL,6pgc}}$
RPE	2	$K_{\mathrm{RPE,ru5p\text{-}D}}$, $K_{\mathrm{RPE,xu5p\text{-}D}}$
RPI	2	$K_{\mathrm{RPI,ru5p\text{-}D}}$, $K_{\mathrm{RPI,r5p}}$
TALA	4	$K_{\mathrm{TALA,g3p}}$, $K_{\mathrm{TALA,s7p}}$, $K_{\mathrm{TALA,e4p}}$, $K_{\mathrm{TALA,f6p}}$
TKT1	4	$K_{\mathrm{TKT1,r5p}}$, $K_{\mathrm{TKT1,xu5p\text{-}D}}$, $K_{\mathrm{TKT1,g3p}}$, $K_{\mathrm{TKT1,s7p}}$
TKT2	4	$K_{\mathrm{TKT2,e4p}}$, $K_{\mathrm{TKT2,xu5p\text{-}D}}$, $K_{\mathrm{TKT2,f6p}}$, $K_{\mathrm{TKT2,g3p}}$

Table 5. 1. PPP reactions and their corresponding kinetic model Km parameters. These reactions correspond to the groups of parameters for which we studied their uncertainty contribution to flux control coefficients.

5.2.2.1. Computation of sensitivity indices

The bounds of the enzyme saturations for all the reactions were initially left unbounded such that we sample them uniformly between 0 and 1, allowing us to consider the full range of kinetic parameter values. We then computed the total effects (St) and first order effects (Si) of the PPP reactions using the Sobol approach (Methods). The St accounted for the contribution to FCC variance of the Km values of a reaction and its interactions with other kinetic parameters. The Si indicates the independent contribution of a reaction's parameters to the variance of the studied model output. Hence, the Si should by definition always be lower or equal to the St value. The closer the Si is to the St, the more the reaction in question uniquely, without interactions, contributes to the variance of the system output variable.

We computed Sobol sensitivity indices for the PPP enzymes for multiple FCCs for a population of 200’000 kinetic models. Since the PPP is composed of twelve enzymatic reactions, we resampled kinetic models to obtain both St and Si for each enzyme (i.e. 2x12 + 1 runs), resulting in a simulation running time of two weeks (Mac Pro, 2.7 GHz 12-Core Intel Xeon E5, 64 GB 1866 MHz DDR3 ECC). The St and Si values for the FCCs of glucose uptake (GLCptspp) with respect to glucose-6-phosphate dehydrogenase 2 (G6PDH2r) (Figure 5.3A) and pyruvate transport (PYRt2rpp) with respect to ribulose-phosphate 3-epimerase (RPE) (Figure 5.3B) summarize a general trend that was observed in FCCs and their Sobol indices. Hence, these FCCs were rather selected for demonstration purpose than for their biological significance. The Si values are always zero for all the parameters studied whereas the St appear to be significant in magnitude (Figure 5.3). This suggested that the PPP reactions’ parameters could interact with the other non-PPP reactions’ parameters to an extent that is considerably more important than their first order effects on the FCCs. Hence, fixing or knowing the parameter values of these PPP reactions would not reduce drastically the uncertainty in the FCCs studied, due to the interactions of the PPP with the rest of the system.

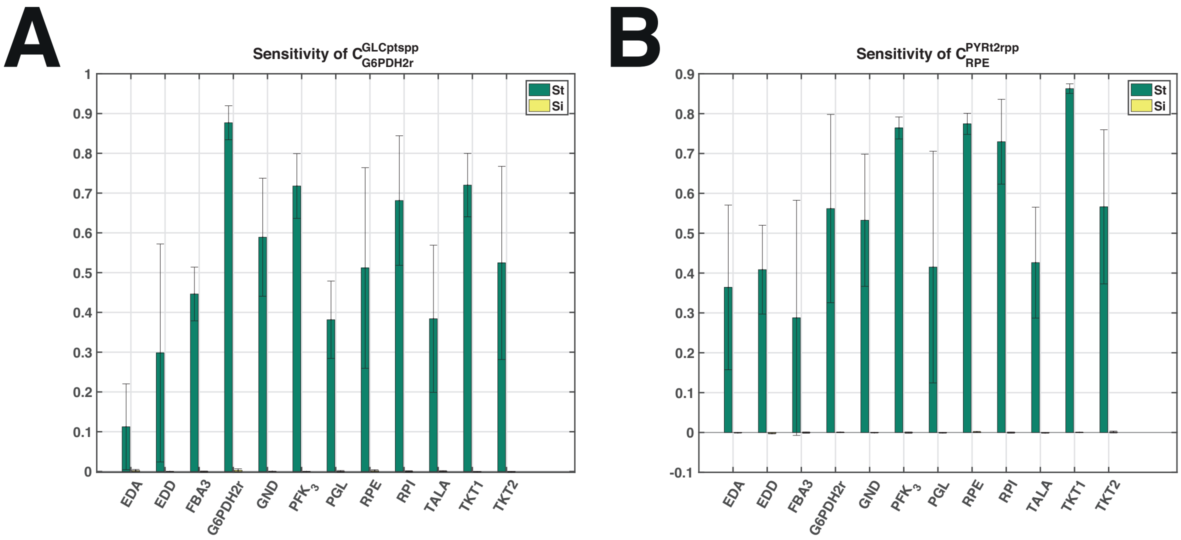


Figure 5. 3. Sobol sensitivity indices for flux control coefficient with respect to PPP enzymes’ saturation levels. Sobol sensitivity indices for St (green) and Si (yellow) of PPP enzymes for (A) glucose uptake control coefficient with respect to glucose-6-phosphate dehydrogenase 2 (G6PDH2r) and (B) pyruvate transporter control coefficient with respect to ribulose-phosphate 3-epimerase (RPE). Enzyme saturation levels were sampled uniformly between 0 and 1. The 200’000 samples were split into three groups, and the

mean and the standard deviation were calculated based on these. The whiskers indicate the standard deviation and the bars report the mean.

5.2.2.2. Kinetic model tightening

As the Si values were zero in the previous example (Figure 5.3), we hypothesized that having the bounds of the non-PPP reactions' parameters unconstrained contributed too significantly uncertainty into the FCCs. We can "tighten" the system by sampling the kinetic parameters for the non-PPP reactions from a smaller range. This reduces the sources of uncertainty stemming from kinetic parameters as the solution space is reduced. The initial results from the previous section were used to compute mean values of the kinetic parameters and we fixed the sampling ranges for the non-PPP reactions to ten percent around this mean. We then repeated the previous experiment to compute St and Si for the tightened kinetic models.

The FCC of GLCptspp with respect to G6PDH2r (Figure 5.4A) does not appear to show significant Si indices, which suggests that the interactions of parameters outside of the PPP dominate and that non of the PPP reactions' kinetic parameters, if known/fixed, can significantly reduce the variance of this FCC. Nevertheless, we notice that the Si values have now become significant for some FCCs (Figure 5.4B). The FCC of PYRt2rpp with respect to RPE appears to be most sensitive to RPE, transketolase (TKT1) and ribose-5-phosphate isomerase (RPI) from the PPP, displaying Si values of 0.26, 0.25 and 0.08, respectively. Hence, fixing the parameters in RPE and TKT1 should reduce the variance of this FCC by up to 50%. However, we noted that the interactions of PPP reactions with other parameters remain significant in terms of decomposition of variance of FCCs. Hence, reducing the size of the kinetic parameter ranges for network parts that are not part of the study scope is an efficient method for artificially reducing the sources of uncertainty, allowing one to focus on determining which parameters, or groups of parameters, from the study scope contribute most to uncertainty. However, it could be beneficial to further study different levels of tightening as multiple different ways and degrees of tightening can be devised.

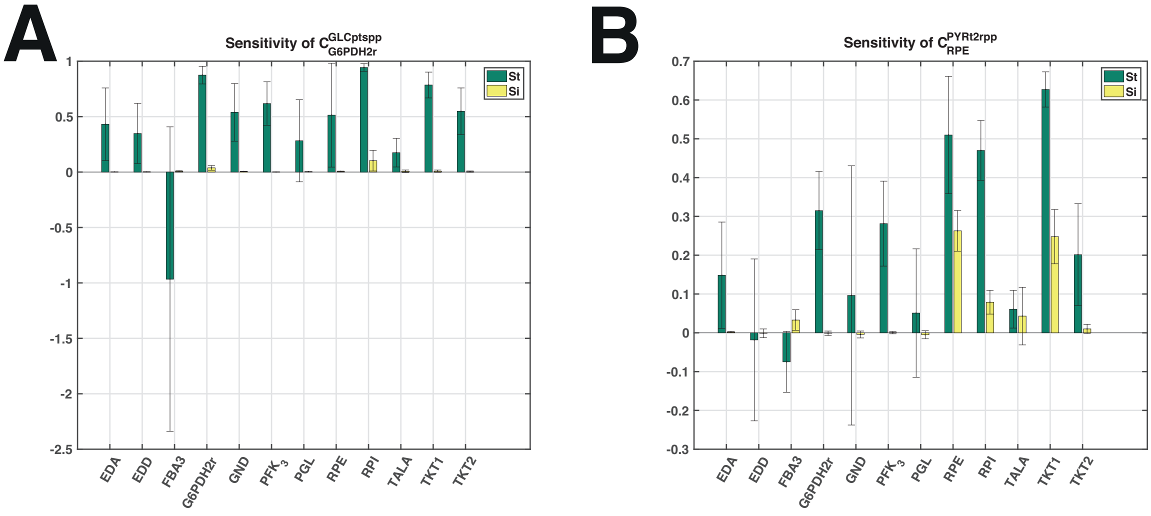


Figure 5. 4. Sobol sensitivity indices for flux control coefficient with respect to PPP enzymes' saturation levels. Sobol sensitivity indices for St (green) and Si (yellow) of PPP enzymes for (A) glucose uptake control coefficient with respect to glucose-6-phosphate dehydrogenase 2 (G6PDH2r) and (B) pyruvate transporter control coefficient with respect to ribulose-phosphate 3-epimerase (RPE). Enzyme saturation levels were sampled uniformly between 0 and 1 for the PPP-reactions. The non-PPP reactions' saturation ranges were anchored to 10% around mid saturation levels determined from the mean value taken over a previous population of kinetic models (further explanations in the main text). The 350'000 samples were split into three groups, and the mean and the standard deviation were calculated based on these. The whiskers indicate the standard deviation and the bars report the mean.

5.2.2.3. Application and validation

The results from the previous section suggested that if we knew the kinetic parameter values of RPE and TKT1 kinetic parameters (Figure 5.4B), we could reduce the variability of the FCC of PYRt2rpp with respect to RPE by around 50%. The initial variance of this FCC was recorded to be 0.27. When we reduced the ranges of the sampled kinetic parameters for RPE and TKT1 to a range of 10% around medium saturation level, rather than sampling uniformly between 0 and 1, we reduced the variance of this FCC to 0.07. This is a reduction in variance of 74%, which is larger than expected based on the Si of RPE and TKT1 combined (Figure 5.4B). The error bars on the Si estimates are relatively large based on the distributions that provide approximations of the Sobol indices. Taking this uncertainty in the Sobol indices into account, the obtained reduction of variance isn't far away from our estimate.

We also ranked the enzymes from PPP for the FCC of ethanol transporter (ETOHtrpp) with respect to transketolase enzyme (TKT1) according to our GMCA workflow (Figure 5.5). Based on the Si values, we could reduce the variability of the FCC of ETOHtrpp with respect to TKT1 by around 68% by fixing transketolases TKT1 and TKT2. The initial variance of this FCC was recorded to be 0.0071 but, when we reduced the ranges of the sampled kinetic parameters for TKT1 and TKT2 to a range of 10% around medium saturation level, we reduced the variance of this FCC to 0.0005. This is a reduction in variance of 93%, which is again larger than expected based on the Si of TKT1 and TKT2 combined (Figure 5.5). This suggests that the Sobol approach appears to well indicate the parameters that contribute to the uncertainty, but, the indices should be studied with care as we note that the St indices have error bars that are sometimes very large (Figure 5.4 and 5.5). The Sobol indices may not entirely converge, which could be due to the complicated nature of the distributions of some of the studied FCCs.

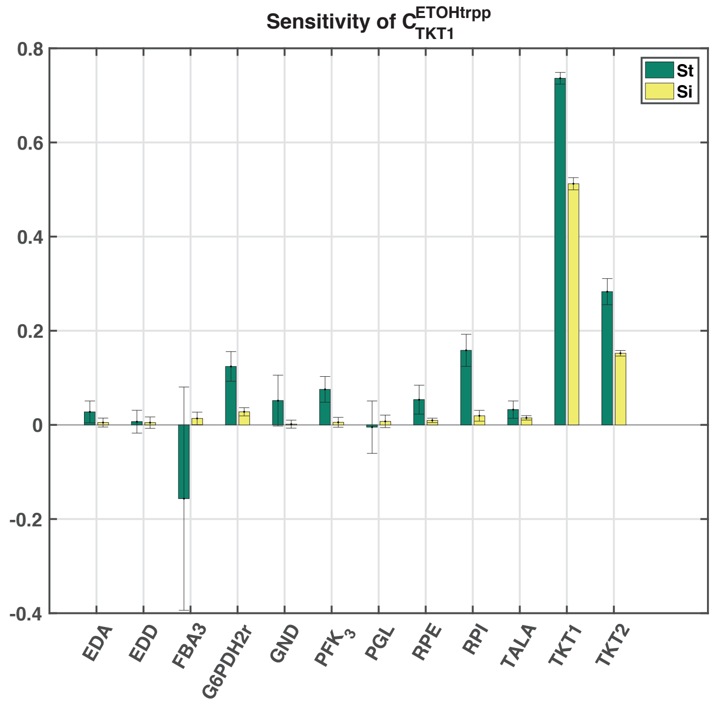


Figure 5. 5. Sobol sensitivity indices for flux control coefficient with respect to PPP enzymes' saturation levels. Sobol sensitivity indices for St (green) and Si (yellow) of PPP enzymes for ethanol transporter (ETOHtrpp) with respect to transketolase (TKT1). Enzyme saturation levels were sampled uniformly between 0 and 1 for the PPP-reactions. The non-PPP reactions' saturation ranges were anchored to 10% around mid saturation levels determined from the mean value taken over a previous population of kinetic models (further explanations in the main text). The 350'000 samples were split into three groups,

and the mean and the standard deviation were calculated based on these. The whiskers indicate the standard deviation and the bars report the mean.

As we discussed in chapter four, the FCCs have distributions that are generally complex, do not behave normally and usually have relatively heavy tails. We computed the kurtosis for the previous run, when fixing TKT1 and TKT2 parameters, that this measure of distribution skewedness does not fully converge for the FCC (Appendix D1). Consequently, even larger sample sizes and better sampling methods should be considered for future studies to ensure convergence of the different statistical modes. Nevertheless, the Si coefficients' error bars are relatively small, indicating their correctness as their implications seem biologically viable (Figure 5.4 and 5.5). However, as expected [129], the St requires more samples than the Si in order to converge and its error bars are considerably larger than for the Si. Different metrics for quantifying the convergence of Sobol indices should be considered.

5.2.3. Future opportunities and limitations

We developed a workflow for applying variance-based sensitivity analysis approaches to large-scale kinetic models of metabolism with an example around PPP. However, this method can be adapted to different study scopes. We could for instance compare how user-defined groups of parameters contribute to the variance of control coefficients. This is what we refer to as a coarse-grain sampling approach as we are grouping multiple parameters. Alternatively, we could focus around comparing how several parameters independently contribute to the variance in kinetic model outputs. This is a higher resolution method that we refer to as fine-grain sampling. However, we must note that this method becomes computationally very expensive as we wish to increase the number of parameters to be studied. Nevertheless, this can be used for experimental design in determining for instance which Km value from a given reaction we should measure in order to maximize variance reduction of FCCs.

We previously pointed towards the main limitation of this method, which is the requirement for very large numbers of samples in order to reach convergence. This is particularly hindering when dealing with very skewed distributions such as the FCCs. The distributions of the FCCs do not generally behave in a normal manner and we occasionally observe outliers that contribute to the distributions having heavy tails. Furthermore, the convergence of statistical modes for FCCs can take large numbers of samples due the complexity of the system and the nature of the data. Consequently, it takes even more samples and computational resources to reach convergence of the Sobol indices. Sensitivity analysis and the convergence of underlying sensitivity indices has been studied in pharmacokinetic [130] and environmental [129] modeling and similar observations have been made. For an environmental model with fifty parameters, around half a million samples can be required for convergence of Sobol indices [129]. However, with technological advances and the increasing availability of computational power, variance-based sensitivity analysis approaches are becoming more accessible for large-scale kinetic models. Alternatively, Kiparissides and coworkers have suggested that derivative-based global sensitivity measures (DGSM) [68] provide a more efficient method for performing sensitivity analysis.

Another area of investigation would be to further study the populations of kinetic parameters to understand how they result in the FCC distributions having such heavy tails. It could be expected from previously done research that several kinetic parameters explain this behavior [49, 66]. Hence, it could be interesting to use machine learning algorithms such as the iSHRUNK workflow [49] that is based around classification and regression tree (CART), to study which kinetic parameters contribute to the complexity of the FCC distributions. The CART approach could reveal which parameters cause this behavior and unravel information about parameter ranges and limits. It would not be surprising if regions of bistability can be found in these systems as they have been observed in biological organisms [131, 132], including *E.coli* [133]. Nevertheless, understanding what parameters cause the FCCs distributions to have heavy tails could further help characterize sources of uncertainty, and it may be advisable in future studies to perform such analysis prior to carrying out variance-based sensitivity analysis. The iSHRUNK workflow could be a useful starting point for performing these studies [49].

5.3. Conclusions

We hereby introduced a workflow that performs variance-based sensitivity analysis on large-scale kinetic models of metabolism. To our knowledge, this type of study has not been carried out on kinetic models that were as large, as most previous studies were performed on systems with several dozens of parameter. We demonstrated how we can complete a sensitivity analysis around the PPP in order to rank the reactions based on their contribution to uncertainty in FCCs and suggested how this method can be adapted to different study scopes. We highlighted that the method can be computationally very expensive but, as computational resources are becoming increasingly available with technological advances, the study can readily be applicable to large systems, like the one demonstrated here. We drew the readers attention to the point that care should be taken when analyzing outputs of kinetic models with variance-based sensitivity analysis, particularly when they behave in a non-normal manner. It may be advisable to study the kinetic model outputs first, prior to engaging into sensitivity analysis. We point out that machine learning approaches, such as iSHRUNK [49], could open up new avenues for unraveling information about parameters that contribute to uncertainty, particularly when having to handle data that is of a complex nature. As an alternative, we also suggest that DGSM [68] could provide more efficiency in performing a sensitivity analysis.

5.4. Materials and methods

5.4.1. MCA, sampling saturations

Kaeser and Burns [34] defined the concentration control coefficients ($\mathrm{C_p^x}$) and the flux control coefficients ($\mathrm{C_p^v}$) as the fractional change of the metabolite concentrations, x, and metabolite fluxes, v, respectively, in response to a fractional change in system parameters p. From the log(linear) formalism, we can derive them as:

$$\mathrm{C_p^x} = -(\mathrm{NVE})^{-1}\mathrm{NV\Pi}$$

$$\mathrm{C_p^v} = \mathrm{EC_p^x} + \Pi$$

where, N is the stoichiometric matrix, V is the diagonal matrix containing the steady-state fluxes, E is the elasticity matrix with respect to metabolites and Π is the matrix of elasticities with respect to parameters.

If we consider a uni-uni reversible Michaelis-Menten enzymatic reaction $\mathrm{S} \leftrightarrow \mathrm{P}$, its thermodynamic displacement, Γ_{i}, is defined as:

$$\Gamma_{\mathrm{i}} = \frac{1}{\mathrm{K}_{\mathrm{eq,i}}} \frac{\mathrm{P}}{\mathrm{S}}$$

where, K_{eq} is the thermodynamic equilibrium constant of the reaction.

The rate expression for this reaction would hence be given by:

$$\nu_{\mathrm{i}} = \nu_{\mathrm{max,i}} \frac{(1-\Gamma_{\mathrm{i}}) \frac{\mathrm{S}}{\mathrm{K}_{\mathrm{mS,i}}}}{1 + \frac{\mathrm{S}}{\mathrm{K}_{\mathrm{mS,i}}} + \frac{\mathrm{P}}{\mathrm{K}_{\mathrm{mP,i}}}} = \nu_{\mathrm{max,i}} \frac{(1-\Gamma_{\mathrm{i}})\, \tilde{\mathrm{S}}_{\mathrm{i}}}{1 + \tilde{\mathrm{S}}_{\mathrm{i}} + \tilde{\mathrm{P}}_{\mathrm{i}}}$$

where, Γ_{i} is the thermodynamic displacement of the reaction i and $\nu_{\mathrm{max,i}}$ is its maximum flux. $\mathrm{K}_{\mathrm{mS,i}}$ and $\mathrm{K}_{\mathrm{mP,i}}$ correspond to the Michaelis-Menten constants of metabolites S and P, respectively. $\tilde{\mathrm{S}}_{\mathrm{i}}$ and $\tilde{\mathrm{P}}_{\mathrm{i}}$ are the metabolite concentrations S and P scaled by their corresponding Michaelis-Menten constants.

The kinetic parameter space is characterized by uniformly sampling the saturation terms of reaction mechanisms [84]. The saturation, σ, is the fraction of a binding site that is occupied by a substrate and is by definition well bounded $\in [0,1]$. We define the saturation of the enzyme of reaction i with respect to S as:

$$\sigma = \frac{\frac{\mathrm{S}}{\mathrm{K}_{\mathrm{mS,i}}}}{1 + \frac{\mathrm{S}}{\mathrm{K}_{\mathrm{mS,i}}}} = \frac{\tilde{\mathrm{S}}_{\mathrm{i}}}{1 + \tilde{\mathrm{S}}_{\mathrm{i}}}$$

We can hence define the scaled concentrations in terms of the sampled saturations as:

$$\tilde{\mathrm{S}}_{\mathrm{i}} = \frac{\sigma}{1-\sigma}$$

As illustrated by Fell and Sauro [134], the elasticities of the reaction with respect to its metabolites directly depend on the scaled concentrations and are given as:

$$\mathrm{E}_{\mathrm{i,S}} = \frac{\partial \ln \nu_{\mathrm{i}}}{\partial \ln x_{\mathrm{S}}} = \frac{1}{1 - \Gamma_{\mathrm{i}}} - \frac{\tilde{\mathrm{S}}_{\mathrm{i}}}{1 + \tilde{\mathrm{S}}_{\mathrm{i}} + \tilde{\mathrm{P}}_{\mathrm{i}}}$$

$$\mathrm{E}_{\mathrm{i,P}} = \frac{\partial \ln \nu_{\mathrm{i}}}{\partial \ln x_{\mathrm{P}}} = -\frac{\Gamma_{\mathrm{i}}}{1 - \Gamma_{\mathrm{i}}} - \frac{\tilde{\mathrm{P}}_{\mathrm{i}}}{1 + \tilde{\mathrm{S}}_{\mathrm{i}} + \tilde{\mathrm{P}}_{\mathrm{i}}}$$

Sampling saturations facilitates the computation of scaled concentrations that can directly be used to populate the elasticy matrix E required for computing the control coeffcients.

5.4.2. GSA, calculating sensitivity indices

Generate matrix A as the base case, consisting of N samples and k input parameters. The input parameters are the sampled σ values. We compute matrix $\mathrm{B_j}$ by fixing parameters of the column j of matrix A into $\mathrm{B_j}$ and we resample the rest for b. Similarly, we compute $\mathrm{C_j}$ by fixing all parameters of A except the column j into it, essentially meaning that we only resample column j.

$$\mathrm{A} = \begin{pmatrix} a_{1,1} & \cdots & a_{1,k} \\ \vdots & \ddots & \vdots \\ a_{N,1} & \cdots & a_{N,k} \end{pmatrix}$$

$$\mathrm{B_j} = \begin{pmatrix} b_{1,1} & \cdots & a_{1,j} & \cdots & b_{1,k} \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ b_{N,1} & \cdots & a_{N,j} & \cdots & b_{N,k} \end{pmatrix}$$

$$\mathrm{C_j} = \begin{pmatrix} a_{1,1} & \cdots & c_{1,j} & \cdots & a_{1,k} \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ a_{N,1} & \cdots & c_{N,j} & \cdots & a_{N,k} \end{pmatrix}$$

The matrices $\mathrm{B_j}$ and $\mathrm{C_j}$ are constructed by resampling the values of b and c, respectively. Sometimes a stable model can not be obtained for a set of parameters (row) from A. We attempted up to 1000 trials for each sample of A in order to generate $\mathrm{B_j}$ and $\mathrm{C_j}$. We evaluate the models to generate output vectors $\mathrm{y_A}$, $\mathrm{y_{B_j}}$ and $\mathrm{y_{C_j}}$, which are the flux control coefficients in this case.

Sensitivity indices are calculated as follows for an output Y:

$$\mathrm{Si} = \frac{\mathrm{V_{Xj}}\left(\mathrm{E_{X\sim j}}(\mathrm{Y}|\mathrm{X_j})\right)}{\mathrm{V(Y)}} = \frac{\frac{\mathrm{y_A y_{Bj}}}{\mathrm{N}} - \mathrm{f_0^2}}{\frac{\mathrm{y_A y_A}}{\mathrm{N}} - \mathrm{f_0^2}}$$

$$\mathrm{St} = 1 - \frac{\mathrm{V_{X\sim j}}\left(\mathrm{E_{Xj}}(\mathrm{Y}|\mathrm{X_{\sim j}})\right)}{\mathrm{V(Y)}} = 1 - \frac{\frac{\mathrm{y_A y_{Cj}}}{\mathrm{N}} - \mathrm{f_0^2}}{\frac{\mathrm{y_A y_A}}{\mathrm{N}} - \mathrm{f_0^2}}$$

where, Si and St are the first order and total effects, respectively. The Si corresponds to the independent contribution of a parameter to the variance of the measurable output Y. The St is the total contributions of a parameter (including its interaction with other parameters) to the variance of the model output Y. Hence, by definition:

$$0 \leq \mathrm{Si} \leq \mathrm{St} \leq 1$$

When Si is equal to St, we can say that the parameter in question is uniquely responsible for the variance of a model output. For full derivations of the sensitivity indices the reader is referred to Sobol's publication [128].

5.5. Appendix D

D1 Figure. Kurtosis convergence plot. Plot of kurtosis convergence for flux control coefficient of ethanol transporter (ETOHtrpp) with respect to transketolase (TKT1). Enzyme saturation levels were sampled uniformly between 0 and 1 for the PPP-reactions. The non-

PPP reactions’, TKT1 and TKT2 saturation ranges were anchored to 10% around mid saturation levels determined from the mean value taken over a previous population of kinetic models (further explanations in the main text).

6. Conclusions and future perspectives

Despite being the most understood cellular process to date, metabolism is still not completely comprehended. Mathematical modeling of metabolism has been proven to provide, when combined with experimental studies, invaluable information about physiologies and cellular behavior [32]. As advances in technology are enabling gene editing and the elaboration of cellular factories that can produce chemicals of interest, constructing mathematical models that provide useful and correct information is essential in guiding these developments [105]. In **Chapter 1**, we established that kinetic modeling of metabolism could elucidate regulatory properties about cells and direct experimentalist in designing insightful investigations. We also identified certain hindering knowledge gaps that we investigated in **Chapters 2-5** in order to develop a workflow for constructing more comprehensive and consistent kinetic models of metabolism. This chapter concludes with the main contributions of this thesis to the field and discusses some future opportunities for research.

6.1. Conclusions

When examining the state of the art of mathematical modeling in **Chapter 1**, we identified that the key issues that underlie the construction of kinetic models are related to two general themes: network topology and handling the various sources of uncertainty. In terms of network topology, we identified that most of the kinetic models constructed to date are built via *ad hoc* protocols and that there are often no systematic justifications nor explanations for their size and content. The other recognized problem is the management of underlying uncertainty in constructing kinetic models that can be categorized into three types: steady state (metabolite concentrations and metabolic fluxes), kinetic mechanisms describing the system and the numerical values of parameters making up kinetic mechanisms.

In **Chapter 2** we addressed the issue of network topology as, to our knowledge, there are no evident and systematic procedures for constructing kinetic models of metabolism for a physiology. We used the redGEM and the lumpGEM algorithms [57, 58] to construct three consistently and modularly reduced stoichiometric models from the *iJO1366* genome-scale model [77] for aerobically grown *E.coli* [61]. These three models were of increasing complexity in terms of network topology and served as basis for building kinetic models using the ORACLE framework [50, 51, 84-86]. We proposed a way for scaling up steady states of the metabolic fluxes and the metabolite concentrations from one model to another and developed a methodology for fixing kinetic parameters between the models in order to preserve equivalency. We performed metabolic control analysis (MCA) around the populations of kinetic models and compared the MCA control coefficients as measurable outputs. We demonstrated that we can systematically reduce genome-scale models to construct kinetic models of different complexity levels for a phenotype that, independent of network complexity, lead to consistent MCA-based metabolic engineering conclusions.

In **Chapter 3** we studied the uncertainty in the metabolite concentrations and the metabolic fluxes as the publications to date generally consider only a steady state when constructing kinetic models. We integrated fluxomics and metabolomcis data for aerobically grown *E.coli* [61] into a consistently reduced model and demonstrated that it was impossible to uniquely determine a steady state due to the underdetermined nature of the system. We built populations of kinetic models around alternative steady states to demonstrate that the selection of a representative steady state can highly impact model-based conclusions. We highlighted that the MCA control coefficients derived for populations of kinetic models were more sensitive to uncertainty in the metabolite concentrations than the metabolic fluxes. A workflow was suggested that allowed the derivation of MCA-based conclusions from kinetic models without neglecting the uncertainty in the fluxes and the concentrations.

In **Chapter 4** we developed some tools for quantifying uncertainty that arises when working with kinetic models. Populations of kinetic models of metabolism are generally constructed in order to derive conclusions about the dynamics of the modeled physiology under uncertainty. However, computational frameworks for building populations of kinetic models do not handle systematically the uncertainty underlying model-based conclusions [55]. Although kinetic models could suggest that modifying the level of certain enzymes would on average increase a flux of interest, this information is incomplete if we do not know with what certainty these model predictions are made. Statistical inference approaches can be used to derive confidence intervals that quantify the level of confidence for which certain model conclusions lie within the intervals. We demonstrated how Bonferroni, exact normal and bootstrapping methodologies can be applied to construct confidence intervals around conclusions derived from populations of kinetic models. There were no considerable differences in the confidence intervals derived with these three methods. The Bonferroni method was notable for its simplicity and for its readiness for estimating the required number of models to attain a level of certainty. We stressed that bootstrapping, despite being complicated computationally and algorithmically, provides certain clear advantages when handling data with highly skewed and/or asymmetric distributions. Regardless of the method used, it is crucial to consider confidence intervals in order to properly quantify uncertainty.

In **Chapter 5** we presented a variance-based global sensitivity analysis (GSA) workflow that can source the uncertainty in populations of large-scale kinetic model outputs to the various input parameters of the system. This appeared to be to date the largest implementation of a sensitivity analysis on kinetic modes of metabolism. We computed sensitivity indices for defined groups of input parameters in order to rank their contribution to the uncertainty of the MCA outputs of the kinetic models. We demonstrated how reactions of the pentose phosphate pathway (PPP) could be ranked in terms of their kinetic parameters' contributions to the uncertainty of MCA flux control coefficients. Information derived from the sensitivity analysis can guide the design of experiments in order to measure/estimate parameters that contribute significantly to uncertainty. This workflow can readily be adapted to different study scopes as the user can define parts of the network for analysis. Admittedly, the workflow is computationally expensive but can become more accessible as computational power, and access to it, improves with technological advances. Additionally, derivative-based sensitivity analysis approaches [67, 135] could provide new opportunities for further developing the current work and reducing computational costs.

Overall, this thesis contributed towards establishing a general workflow (Figure 6.1) for constructing comprehensive study-specific kinetic models of metabolism and proposed approaches and tools for coping with the underlying uncertainties. The workflow was developed with studies of *E.coli* metabolism but it could be applied to other organisms for carrying out similar work. Such studies can provide valuable information about the metabolism of different living entities and deliver insight into designing experimental studies and making metabolic engineering decisions. This computational workflow that combines data from various sources could be coupled with further experimental studies to validate and enhance the undertaken analysis, thereby completing and re-iterating the systems biology cycle.

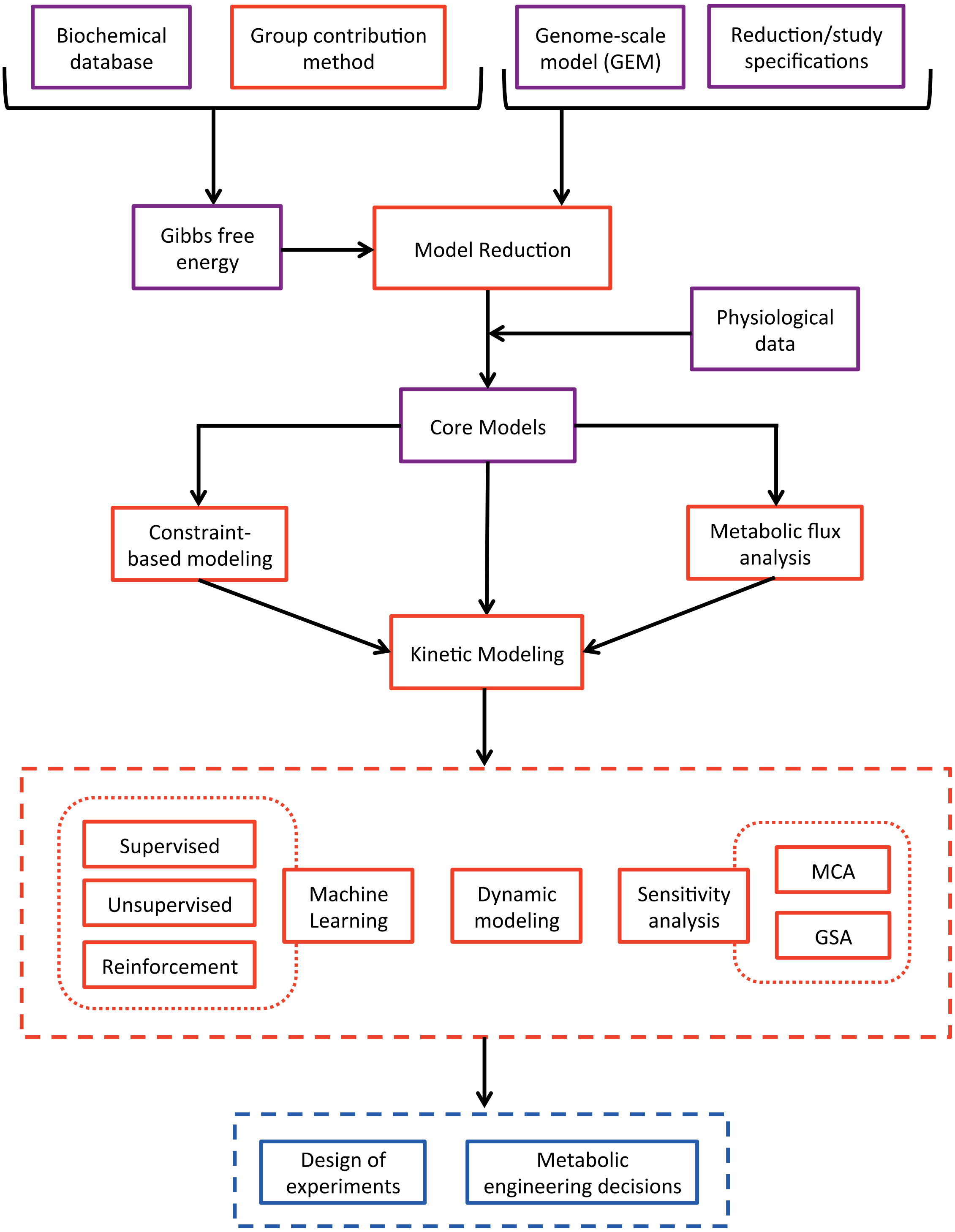


Figure 6. 1. Workflow for constructing consistent kinetic models of metabolism.

6.2. Future perspectives

As technological advances are enabling the characterization of various physiological properties of diverse living organisms, the future of systems biology promises plenty of opportunities for incorporating these data into mathematical models. With the increasing availability of metabolomics data [61, 106, 136], these can be integrated into models to reduce uncertainty in reaction directionalities via thermodynamic constraints [13-16], as we have done in **Chapter 2** and **Chapter 3**. As metabolite concentrations and their change with time can be measured more precisely, experiments could explain metabolic interactions and provide detailed accounts on cellular regulation for some enzymatic reactions [137, 138]. Such advances could facilitate the incorporation of more detailed mathematical descriptions of kinetic mechanisms into models and thus improve their accuracy and reliability.

As data about kinetics become increasingly available [93, 94], the inclusion of Km data into kinetic models will reduce uncertainty in these parameters. Since we sample Km values between defined ranges of enzyme saturations, we could also use Km data to constrain metabolite concentrations in stoichiometric models, particularly if experimental measurements of their concentration are not available. Additionally, as proteomic and catalytic turnover rate data are becoming more abundant [139, 140], the estimation of Vmax values could allow further constraining of allowable fluxes across enzymatic reactions. Inclusion of such data within the thermodynamic-based framework would further reduce uncertainty when selecting representative steady states for constructing kinetic models of the physiology [25]. Because reaction rates are not *di per se* measurable, experiments employing 13C tracers [106-108] could provide estimation of the fluxes across reactions, thus helping to characterize steady states.

Besides expecting an exponential increase in available physiological data, new approaches may be required in order to cope with and analyze vast quantities of data and to derive meaningful conclusions. Certain algorithms that employ machine learning (Figure 6.1), like the iSHRUNK workflow [49], have already been applied successfully to study populations of kinetic models and could provide useful guidance in the future. Exploration of the various omics data with machine learning algorithms could offer new insight on regulatory mechanisms and enhance understanding about system dynamics [141, 142]. Until more precise data about cellular dynamics and kinetic mechanisms become available, kinetic modeling will heavily rely on Monte Carlo sampling approaches [143] in order to generate values for unknown parameters and build populations of models. Nevertheless, the work from this thesis can be applied for the development of personalized medicine, targeted metabolic engineering strategies and models combining metabolism with signaling, as we discuss below.

6.2.1. Personalized medicine

Many people are taking medication that will not benefit them [144]. This is because every individual can respond differently to the ingested medication. Genetics and environmental factors can vary strongly between people and hence their body will react differently. However, medication is generally designed based on average observations and response of patients to certain clinical trials. The problem with this approach is that it assumes that people will respond similarly to administered pharmaceuticals. Hence, it is important to consider patients individually or in groups defined by certain physiological traits when designing treatments. Systems biology of metabolism can provide significant insight into such developments [145, 146] and the workflow (Figure 6.1) devised in this thesis can be applied.

Genome-scale models of human metabolism have been published [147-149] and can serve as chassis for constructing kinetic models [150]. The lumpGEM and redGEM algorithms [57, 58] can be used for model reduction based on study specifications. The gut microbiome changes considerably between individuals and could be held responsible for differences in response to medication in humans [144]. Technologies have enabled the comprehensive metabolite profiling of blood [151] and the gut microbiome [152, 153]. Data about the gut microbiome could be integrated into reduced stoichiometric models in order to mirror alternative cellular environments/media of different patients. These stoichiometric models can be used to construct personalized kinetic models using our workflow (Figure 6.1) and could unravel metabolic signatures and variations in the regulatory mechanisms across individuals. Such knowledge derived from the kinetic models can help pharmaceutical companies and the health care institutions to identify biomarkers and hence, to devise personalized and precision medicine.

6.2.2. Targeted metabolic engineering strategies

Metabolic engineering is the targeted improvement of physiological traits of cells via modification of certain biochemical reactions and/or the incorporation of new ones via genome editing. The former approach requires information about which enzymes should be modified in order to achieve certain flux distribution and, MCA has been suggested as a computational approach that could provide such knowhow [154]. The latter method requires knowledge about reactions or pathways that should be added to the system in order to achieve a desirable physiology. To this end, computational frameworks that propose biochemical and novel reactions/pathways that connect study-specific pairs of metabolites [155, 156] can be combined with pathway evaluation methods to suggest such metabolic engineering strategies [157]. Whichever methodology is used, if not a hybrid of both, information about the regulatory properties of the system are desirable in order to further optimize the metabolic engineering strategies [158].

The workflow (Figure 6.1) can be applied to these systems in order to derive systematically reduced kinetic models for the physiology in question. We could also construct kinetic models around certain competing physiologies in order to compare how their metabolic control patterns differ based on MCA. Mixed-integer linear programming (MILP) optimization techniques [110] can be applied to MCA outputs to find optimum strategies for these physiologies and, coupled with expert knowledge, they can be compared and evaluated. Such analysis could be carried out with the same organism for alternative physiologies or even, it could be extended to comparing how different organisms would fair as cell factories. *E.coli* and yeast provide robust organisms for designing cell factories as they are the best characterized in terms of available physiological data [159, 160]. Nevertheless, as genome-scale metabolic models of multiple organisms exist [161] and are being published, numerous opportunities lie out there for testing how other organisms could perform as cell factories for metabolic engineering using our workflow and, the above suggestions.

6.2.3. Models combining metabolism with signaling

Cellular function can be studied at different levels: metabolism, transcriptional regulation and signaling. These three processes have generally been studied independently even though it is known that there is interplay between them [162]. Limitations and advances in integrating information from these different levels together into mathematical models have been discussed extensively in reviews [31, 163, 164]. Signaling is known to be in control of cellular growth and metabolism [165]. However, combining data about signaling and metabolism into a kinetic model remains challenging. Signaling is generally modeled using discrete Boolean state variables that are difficult to translate into continuous time-resolved predictions [166]. Metabolism is usually modeled as a continuous process that can be described by ordinary differential equations (ODEs). Despite these differences in modeling approaches employed in studies of signaling and metabolism, new methodologies [166, 167] could be applied to integrate these two processes into kinetic models.

For instance, insulin plays an important role in controlling metabolism in humans [168]. The workflow (Figure 6.1) developed in this thesis could serve as basis for deriving kinetic models of human metabolism from published genome-scale models [147-149]. These kinetic models provide the system of ODEs describing cellular metabolism. Information about metabolic actions of insulin in humans [169] could then be integrated into these kinetic models via techniques for transforming Boolean models into continues ones [167]. As experimental technologies have significantly advanced, optical intracellular tracking of proteins [170] could further elucidate details about interactions between signaling and metabolism. The construction of kinetic models could be coupled with such experimental studies in order to comprehensively integrate signaling and metabolism into mathematical models.

Bibliography

1. Feist AM, Herrgård MJ, Thiele I, Reed JL, Palsson BØ. Reconstruction of biochemical networks in microorganisms. Nature Reviews Microbiology. 2009;7(2):129.

2. Thiele I, Palsson BØ. A protocol for generating a high-quality genome-scale metabolic reconstruction. Nature protocols. 2010;5(1):93.

3. Lewis NE, Nagarajan H, Palsson BO. Constraining the metabolic genotype–phenotype relationship using a phylogeny of in silico methods. Nature Reviews Microbiology. 2012;10(4):291.

4. Varma A, Palsson BO. Stoichiometric flux balance models quantitatively predict growth and metabolic by-product secretion in wild-type Escherichia coli W3110. Applied and environmental microbiology. 1994;60(10):3724-31.

5. Edwards J, Covert M, Palsson B. Metabolic modelling of microbes: the flux-balance approach. Environ Microbiol. 2002;4:133 - 40. PubMed PMID: doi:10.1046/j.1462-2920.2002.00282.x.

6. Orth JD, Thiele I, Palsson BO. What is flux balance analysis? Nat Biotech. 2010;28(3):245-8. doi: http://www.nature.com/nbt/journal/v28/n3/abs/nbt.1614.html - supplementary-information.

7. Schuetz R, Kuepfer L, Sauer U. Systematic evaluation of objective functions for predicting intracellular fluxes in Escherichia coli2007 2007-01-01 00:00:00.

8. Becker SA, Feist AM, Mo ML, Hannum G, Palsson BO, Herrgard MJ. Quantitative prediction of cellular metabolism with constraint-based models: the COBRA Toolbox. Nat Protocols. 2007;2:727-38. doi: 10.1038/nprot.2007.99.

9. Schellenberger J, Que R, Fleming RMT, Thiele I, Orth JD, Feist AM, et al. Quantitative prediction of cellular metabolism with constraint-based models: the COBRA Toolbox v2.0. Nat Protocols. 2011;6:1290-307. doi: 10.1038/nprot.2011.308.

10. Beard DA, Liang S-d, Qian H. Energy balance for analysis of complex metabolic networks. Biophysical journal. 2002;83(1):79-86.

11. Kümmel A, Panke S, Heinemann M. Putative regulatory sites unraveled by network-embedded thermodynamic analysis of metabolome data. Molecular systems biology. 2006;2(1).

12. Zamboni N, Kümmel A, Heinemann M. anNET: a tool for network-embedded thermodynamic analysis of quantitative metabolome data. BMC bioinformatics. 2008;9(1):199.

13. Henry CS, Broadbelt LJ, Hatzimanikatis V. Thermodynamics-Based Metabolic Flux Analysis. Biophysical Journal. 2007;92:1792 - 805. doi: http://dx.doi.org/10.1529/biophysj.106.093138.

14. Soh KC, Hatzimanikatis V. Constraining the Flux Space Using Thermodynamics and Integration of Metabolomics Data. Metabolic Flux Analysis: Springer; 2014. p. 49-63.

15. Salvy P, Fengos G, Ataman M, Pathier T, Soh KC, Hatzimanikatis V. pyTFA and matTFA: A Python package and a Matlab toolbox for Thermodynamics-based Flux Analysis. Bioinformatics. 2018;1:3.

16. Ataman M, Hatzimanikatis V. Heading in the right direction: thermodynamics-based network analysis and pathway engineering. Current opinion in biotechnology. 2015;36:176-82.

17. Mavrovouniotis ML. Group contributions for estimating standard gibbs energies of formation of biochemical compounds in aqueous solution. Biotechnology and Bioengineering. 1990;36(10):1070-82. doi: 10.1002/bit.260361013.

18. Mavrovouniotis M. Estimation of standard Gibbs energy changes of biotransformations. J Biol Chem. 1991;266(22):14440 - 5.

19. Lerman JA, Hyduke DR, Latif H, Portnoy VA, Lewis NE, Orth JD, et al. In silico method for modelling metabolism and gene product expression at genome scale. Nature communications. 2012;3:929.

20. O'brien EJ, Lerman JA, Chang RL, Hyduke DR, Palsson BØ. Genome-scale models of metabolism and gene expression extend and refine growth phenotype prediction. Molecular systems biology. 2013;9(1):693.

21. Thiele I, Fleming RM, Que R, Bordbar A, Diep D, Palsson BO. Multiscale modeling of metabolism and macromolecular synthesis in E. coli and its application to the evolution of codon usage. PloS one. 2012;7(9):e45635.

22. Liu JK, O'Brien EJ, Lerman JA, Zengler K, Palsson BO, Feist AM. Reconstruction and modeling protein translocation and compartmentalization in Escherichia coli at the genome-scale. BMC systems biology. 2014;8(1):110.

23. Yang L, Yurkovich JT, Lloyd CJ, Ebrahim A, Saunders MA, Palsson BO. Principles of proteome allocation are revealed using proteomic data and genome-scale models. Scientific reports. 2016;6:36734.

24. Ebrahim A, Brunk E, Tan J, O'brien EJ, Kim D, Szubin R, et al. Multi-omic data integration enables discovery of hidden biological regularities. Nature communications. 2016;7:13091.

25. Sánchez BJ, Zhang C, Nilsson A, Lahtvee PJ, Kerkhoven EJ, Nielsen J. Improving the phenotype predictions of a yeast genome-scale metabolic model by incorporating enzymatic constraints. Molecular systems biology. 2017;13(8):935.

26. Miskovic L, Hatzimanikatis V. Production of biofuels and biochemicals: in need of an {ORACLE}. Trends in Biotechnology. 2010;28:391 - 7. doi: http://dx.doi.org/10.1016/j.tibtech.2010.05.003.

27. Mahadevan R, Edwards JS, Doyle III FJ. Dynamic flux balance analysis of diauxic growth in Escherichia coli. Biophysical journal. 2002;83(3):1331-40.

28. Covert MW, Xiao N, Chen TJ, Karr JR. Integrating metabolic, transcriptional regulatory and signal transduction models in Escherichia coli. Bioinformatics. 2008;24(18):2044-50.

29. Wu H, Von Kamp A, Leoncikas V, Mori W, Sahin N, Gevorgyan A, et al. MUFINS: multi-formalism interaction network simulator. NPJ systems biology and applications. 2016;2:16032.

30. Luo RY, Liao S, Tao GY, Li YY, Zeng S, Li YX, et al. Dynamic analysis of optimality in myocardial energy metabolism under normal and ischemic conditions. Molecular systems biology. 2006;2(1).

31. Chiappino-Pepe A, Pandey V, Ataman M, Hatzimanikatis V. Integration of metabolic, regulatory and signaling networks towards analysis of perturbation and dynamic responses. Current Opinion in Systems Biology. 2017;2:59-66.

32. Nielsen J. Systems Biology of Metabolism. Annual Review of Biochemistry. 2017;86(1):245-75. doi: 10.1146/annurev-biochem-061516-044757.

33. Almquist J, Cvijovic M, Hatzimanikatis V, Nielsen J, Jirstrand M. Kinetic models in industrial biotechnology – Improving cell factory performance. Metabolic Engineering. 2014;24:38 - 60. doi: http://dx.doi.org/10.1016/j.ymben.2014.03.007.

34. Kacser Ha, Burns J, editors. The control of flux. Symp Soc Exp Biol; 1973.

35. Heinrich R, Rapoport TA. A Linear Steady-State Treatment of Enzymatic Chains. European Journal of Biochemistry. 1974;42(1):89-95. doi: 10.1111/j.1432-1033.1974.tb03318.x.

36. Heinrich R, Schuster S. Introduction. The Regulation of Cellular Systems: Springer US; 1996. p. 1-8.

37. Hatzimanikatis V. Nonlinear Metabolic Control Analysis. Metabolic Engineering. 1999;1(1):75-87. doi: http://dx.doi.org/10.1006/mben.1998.0108.

38. Bakker BM, Westerhoff HV, Opperdoes FR, Michels PA. Metabolic control analysis of glycolysis in trypanosomes as an approach to improve selectivity and effectiveness of drugs. Molecular and biochemical parasitology. 2000;106(1):1-10.

39. Chassagnole C, Noisommit-Rizzi N, Schmid JW, Mauch K, Reuss M. Dynamic modeling of the central carbon metabolism of Escherichia coli. Biotechnology and Bioengineering. 2002;79(1):53-73. doi: 10.1002/bit.10288.

40. Hoefnagel MH, Starrenburg MJ, Martens DE, Hugenholtz J, Kleerebezem M, Van Swam II, et al. Metabolic engineering of lactic acid bacteria, the combined approach: kinetic modelling, metabolic control and experimental analysis. Microbiology. 2002;148(4):1003-13.

41. Cintolesi A, Clomburg JM, Rigou V, Zygourakis K, Gonzalez R. Quantitative analysis of the fermentative metabolism of glycerol in Escherichia coli. Biotechnology and bioengineering. 2012;109(1):187-98.

42. Bakker BM, Michels PA, Opperdoes FR, Westerhoff HV. What controls glycolysis in bloodstream form Trypanosoma brucei? Journal of Biological Chemistry. 1999;274(21):14551-9.

43. Jaqaman K, Danuser G. Linking data to models: data regression. Nature Reviews Molecular Cell Biology. 2006;7(11):813.

44. Chen P-W, Theisen MK, Liao JC. Metabolic systems modeling for cell factories improvement. Current Opinion in Biotechnology. 2017;46:114-9.

45. Khodayari A, Maranas CD. A genome-scale Escherichia coli kinetic metabolic model k-ecoli457 satisfying flux data for multiple mutant strains. Nature Communications. 2016;7.

46. Rivera JGL, Theisen MK, Chen P-W, Liao JC. Kinetically accessible yield (KAY) for redirection of metabolism to produce exo-metabolites. Metabolic engineering. 2017;41:144-51.

47. Tran LM, Rizk ML, Liao JC. Ensemble Modeling of Metabolic Networks. Biophysical Journal. 2008;95(12):5606-17. doi: https://doi.org/10.1529/biophysj.108.135442.

48. Andreozzi S, Chakrabarti A, Soh KC, Burgard A, Yang TH, Van Dien S, et al. Identification of metabolic engineering targets for the enhancement of 1,4-butanediol production in recombinant E. coli using large-scale kinetic models. Metabolic Engineering. 2016;35:148-59. doi: 10.1016/j.ymben.2016.01.009. PubMed PMID: 26855240.

49. Andreozzi S, Miskovic L, Hatzimanikatis V. iSCHRUNK–In Silico Approach to Characterization and Reduction of Uncertainty in the Kinetic Models of Genome-scale Metabolic Networks. Metabolic engineering. 2016;33:158-68.

50. Chakrabarti A, Miskovic L, Soh KC, Hatzimanikatis V. Towards kinetic modeling of genome-scale metabolic networks without sacrificing stoichiometric, thermodynamic and

physiological constraints. Biotechnol J. 2013;8(9):1043-57. Epub 2013/07/23. doi: 10.1002/biot.201300091. PubMed PMID: 23868566.

51. Miskovic L, Hatzimanikatis V. Production of biofuels and biochemicals: in need of an ORACLE. Trends in biotechnology. 2010;28(8):391-7.

52. Miskovic L, Hatzimanikatis V. Modelling of uncertainties in biochemical reactions. Biotechnology and Bioengineering. 2011;108:413-23.

53. Soh KS, Miskovic L, Hatzimanikatis V. From network models to network responses: integration of thermodynamic and kinetic properties of yeast genome-scale metabolic networks. FEMS Yeast Research. 2012;12:129-43.

54. Miskovic L, Tokic M, Fengos G, Hatzimanikatis V. Rites of passage: requirements and standards for building kinetic models of metabolic phenotypes. Current Opinion in Biotechnology. 2015;36:1-8.

55. Saa PA, Nielsen LK. Formulation, construction and analysis of kinetic models of metabolism: A review of modelling frameworks. Biotechnology advances. 2017.

56. Teusink B, Passarge J, Reijenga CA, Esgalhado E, van der Weijden CC, Schepper M, et al. Can yeast glycolysis be understood in terms of in vitro kinetics of the constituent enzymes? Testing biochemistry. European Journal of Biochemistry. 2000;267:5313–29. doi: 10.1046/j.1432-1327.2000.01527.x.

57. Ataman M, Gardiol DFH, Fengos G, Hatzimanikatis V. redGEM: Systematic reduction and analysis of genome-scale metabolic reconstructions for development of consistent core metabolic models. PLoS computational biology. 2017;13(7):e1005444.

58. Ataman M, Hatzimanikatis V. lumpGEM: Systematic generation of subnetworks and elementally balanced lumped reactions for the biosynthesis of target metabolites. PLoS computational biology. 2017;13(7):e1005513.

59. Erdrich P, Steuer R, Klamt S. An algorithm for the reduction of genome-scale metabolic network models to meaningful core models. BMC systems biology. 2015;9(1):48.

60. Palsson BØ, Lee I-D. Model complexity has a significant effect on the numerical value and interpretation of metabolic sensitivity coefficients. 1993.

61. McCloskey D, Gangoiti JA, King ZA, Naviaux RK, Barshop BA, Palsson BO, et al. A model-driven quantitative metabolomics analysis of aerobic and anaerobic metabolism in E. coli K-12 MG1655 that is biochemically and thermodynamically consistent. Biotechnology and bioengineering. 2014;111(4):803-15.

62. Park JO, Rubin SA, Xu Y-F, Amador-Noguez D, Fan J, Shlomi T, et al. Metabolite concentrations, fluxes and free energies imply efficient enzyme usage. Nature chemical biology. 2016;12(7):482-9.

63. Soh KC, Hatzimanikatis V. Computational Studies on Cellular Bioenergetics. 2013.

64. Miskovic L, Alff-Tuomala S, Soh KC, Barth D, Salusjärvi L, Pitkänen J-P, et al. A design–build–test cycle using modeling and experiments reveals interdependencies between upper glycolysis and xylose uptake in recombinant S. cerevisiae and improves predictive capabilities of large-scale kinetic models. Biotechnology for biofuels. 2017;10(1):166.

65. Saa P, Nielsen LK. A general framework for thermodynamically consistent parameterization and efficient sampling of enzymatic reactions. PLoS computational biology. 2015;11(4):e1004195.

66. Gutenkunst RN, Waterfall JJ, Casey FP, Brown KS, Myers CR, Sethna JP. Universally sloppy parameter sensitivities in systems biology models. PLoS computational biology. 2007;3(10):e189.

67. Kiparissides A, Georgakis C, Mantalaris A, Pistikopoulos EN. Design of In Silico Experiments as a Tool for Nonlinear Sensitivity Analysis of Knowledge-Driven Models. Industrial & Engineering Chemistry Research. 2014;53(18):7517-25.

68. Kiparissides A, Kucherenko S, Mantalaris A, Pistikopoulos E. Global sensitivity analysis challenges in biological systems modeling. Industrial & Engineering Chemistry Research. 2009;48(15):7168-80.

69. Saltelli A, Ratto M, Andres T, Campolongo F, Cariboni J, Gatelli D, et al. Global sensitivity analysis: the primer: John Wiley & Sons; 2008.

70. Davidi D, Noor E, Liebermeister W, Bar-Even A, Flamholz A, Tummler K, et al. Global characterization of in vivo enzyme catalytic rates and their correspondence to in vitro k(cat) measurements. Proceedings of the National Academy of Sciences of the United States of America. 2016;113(12):3401-6. doi: 10.1073/pnas.1514240113. PubMed PMID: PMC4812741.

71. Saa PA, Nielsen LK. Construction of feasible and accurate kinetic models of metabolism: A Bayesian approach. Scientific reports. 2016;6:29635.

72. Khodayari A, Maranas CD. A genome-scale Escherichia coli kinetic metabolic model k-ecoli457 satisfying flux data for multiple mutant strains. Nature communications. 2016;7:13806.

73. Khodayari A, Zomorrodi AR, Liao JC, Maranas CD. A kinetic model of Escherichia coli core metabolism satisfying multiple sets of mutant flux data. Metabolic engineering. 2014;25:50-62.

74. Teleki A, Rahnert M, Bungart O, Gann B, Ochrombel I, Takors R. Robust identification of metabolic control for microbial l-methionine production following an easy-to-use puristic approach. Metabolic engineering. 2017;41:159-72.

75. Millard P, Smallbone K, Mendes P. Metabolic regulation is sufficient for global and robust coordination of glucose uptake, catabolism, energy production and growth in Escherichia coli. PLoS computational biology. 2017;13(2):e1005396.

76. Palsson BO, Lee I-D. Model complexity has a significant effect on the numerical value and interpretation of metabolic sensitivity coefficients. Journal of theoretical biology. 1993;161(3):299-315.

77. Orth JD, Conrad TM, Na J, Lerman JA, Nam H, Feist AM, et al. A comprehensive genome-scale reconstruction of Escherichia coli metabolism—2011. Molecular systems biology. 2011;7(1):535.

78. Neidhardt FC, Ingraham JL, Schaechter M. Physiology of the bacterial cell: a molecular approach: Sinauer Associates Sunderland, MA; 1990.

79. Cooper R. Metabolism of methylglyoxal in microorganisms. Annual Reviews in Microbiology. 1984;38(1):49-68.

80. Nelson DL, Kennedy EP. Transport of magnesium by a repressible and a nonrepressible system in Escherichia coli. Proceedings of the National Academy of Sciences. 1972;69(5):1091-3.

81. Rosenberg H, Gerdes R, Chegwidden K. Two systems for the uptake of phosphate in Escherichia coli. Journal of bacteriology. 1977;131(2):505-11.

82. Kumble KD, Ahn K, Kornberg A. Phosphohistidyl active sites in polyphosphate kinase of Escherichia coli. Proceedings of the National Academy of Sciences. 1996;93(25):14391-5.

83. Henry CS, Broadbelt LJ, Hatzimanikatis V. Thermodynamics-based metabolic flux analysis. Biophysical journal. 2007;92(5):1792-805.

84. Wang L, Birol I, Hatzimanikatis V. Metabolic Control Analysis under Uncertainty: Framework Development and Case Studies. Biophysical Journal. 2004;87:3750-63.

85. Wang LQ, Hatzimanikatis V. Metabolic engineering under uncertainty - II: Analysis of yeast metabolism. Metabolic Engineering. 2006;8(2):142-59. doi: Doi 10.1016/J.Yinben.2005.11.002. PubMed PMID: ISI:000236053000006.

86. Wang LQ, Hatzimanikatis V. Metabolic engineering under uncertainty. I: Framework development. Metabolic Engineering. 2006;8(2):133-41. doi: Doi 10.1016/J.Ymben.2005.11.003. PubMed PMID: ISI:000236053000005.

87. Heinrich R, Schuster S. The regulation of cellular systems. New York ; London: Chapman & Hall; 1996.

88. Kacser H, Burns J, editors. The control of flux. Symp Soc Exp Biol; 1973.

89. Stephanopoulos G, Vallino JJ. Network rigidity and metabolic engineering in metabolite overproduction. Science. 1991;252(5013):1675-81.

90. Bailey JE. Toward a science of metabolic engineering. Science. 1991;252(5013):1668-75.

91. Jolliffe I. Principal Component Analysis. Wiley StatsRef: Statistics Reference Online: John Wiley & Sons, Ltd; 2014.

92. Segel IH. Enzyme Kinetics. 1975.

93. Schomburg I, Chang A, Placzek S, Sohngen C, Rother M, Lang M, et al. BRENDA in 2013: integrated reactions, kinetic data, enzyme function data, improved disease classification: new options and contents in BRENDA. Nucleic Acids Res. 2013;41(Database issue):D764-72. Epub 2012/12/04. doi: 10.1093/nar/gks1049. PubMed PMID: 23203881.

94. Wittig U, Kania R, Golebiewski M, Rey M, Shi L, Jong L, et al. SABIO-RK-database for biochemical reaction kinetics. Nucleic Acids Res. 2012;40(D1):D790-D6. doi: Doi 10.1093/Nar/Gkr1046. PubMed PMID: ISI:000298601300118.

95. Hofmeyr J, Cornish-Bowden A. The reversible Hill equation: how to incorporate cooperative enzymes into metabolic models. Comput Appl Biosci. 1997;13:377-85.

96. Liebermeister W, Klipp E. Bringing metabolic networks to life: convenience rate law and thermodynamic constraints. Theoretical Biology and Medical Modeling. 2006;3(41). doi: doi:10.1186/1742-4682-3-41.

97. Hatzimanikatis V, Bailey JE. MCA has more to say. Journal of Theoretical Biology. 1996;182(3):233-42.

98. Kim TY, Sohn SB, Kim YB, Kim WJ, Lee SY. Recent advances in reconstruction and applications of genome-scale metabolic models. Current opinion in biotechnology. 2012;23(4):617-23.

99. Alper H, Stephanopoulos G. Engineering for biofuels: exploiting innate microbial capacity or importing biosynthetic potential? Nature Reviews Microbiology. 2009;7:715. doi: 10.1038/nrmicro2186.

100. Li G, Wang J-b, Reetz MT. Biocatalysts for the pharmaceutical industry created by structure-guided directed evolution of stereoselective enzymes. Bioorganic & medicinal chemistry. 2017.

101. Blazeck J, Alper H. Systems metabolic engineering: Genome-scale models and beyond. Biotechnol J. 2010;5(7):647-59.

102. Zamboni N, Fendt S-M, Ruhl M, Sauer U. 13C-based metabolic flux analysis. Nat Protocols. 2009;4(6):878-92. doi: http://www.nature.com/nprot/journal/v4/n6/suppinfo/nprot.2009.58_S1.html.

103. Ebrahim A, Brunk E, Tan J, O'brien EJ, Kim D, Szubin R, et al. Multi-omic data integration enables discovery of hidden biological regularities. Nature Communications. 2016;7.

104. Schuetz R, Kuepfer L, Sauer U. Systematic evaluation of objective functions for predicting intracellular fluxes in Escherichia coli. Molecular Systems Biology. 2007;3:119. doi: 10.1038/msb4100162. PubMed PMID: PMC1949037.

105. Almquist J, Cvijovic M, Hatzimanikatis V, Nielsen J, Jirstrand M. Kinetic models in industrial biotechnology–improving cell factory performance. Metabolic engineering. 2014;24:38-60.

106. Toya Y, Ishii N, Nakahigashi K, Hirasawa T, Soga T, Tomita M, et al. 13C-metabolic flux analysis for batch culture of Escherichia coli and its pyk and pgi gene knockout mutants based on mass isotopomer distribution of intracellular metabolites. Biotechnology progress. 2010;26(4):975-92.

107. Crown SB, Long CP, Antoniewicz MR. Integrated 13 C-metabolic flux analysis of 14 parallel labeling experiments in Escherichia coli. Metabolic engineering. 2015;28:151-8.

108. Fong SS, Nanchen A, Palsson BO, Sauer U. Latent pathway activation and increased pathway capacity enable Escherichia coli adaptation to loss of key metabolic enzymes. Journal of Biological Chemistry. 2006;281(12):8024-33.

109. Kaufman DE, Smith RL. Direction choice for accelerated convergence in hit-and-run sampling. Operations Research. 1998;46(1):84-95.

110. Hatzimanikatis V, Floudas CA, Bailey JE. Analysis and design of metabolic reaction networks via mixed-integer linear optimization. AIChE Journal. 1996;42(5):1277-92.

111. Reder C. Metabolic control theory: a structural approach. Journal of theoretical biology. 1988;135(2):175-201.

112. Hahn GJ, Shapiro SS. Statistical Models in Engineering: Wiley; 1994.

113. Steuer R, Gross T, Selbig J, Blasius B. Structural kinetic modeling of metabolic networks. Proceedings of the National Academy of Sciences. 2006;103(32):11868-73.

114. Kass RE, Caffo BS, Davidian M, Meng X-L, Yu B, Reid N. Ten simple rules for effective statistical practice. PLoS computational biology. 2016;12(6):e1004961.

115. Chakrabarti A, Miskovic L, Soh KC, Hatzimanikatis V. Towards kinetic modeling of genome-scale metabolic networks without sacrificing stoichiometric, thermodynamic and physiological constraints. Biotechnol J. 2013;8(9):1043-57. Epub 2013/07/23. doi: 10.1002/biot.201300091. PubMed PMID: 23868566.

116. Hameri T, Fengos G, Ataman M, Miskovic L, Hatzimanikatis V. Kinetic models of metabolism that consider alternative steady-state solutions of intracellular fluxes and concentrations. Metabolic Engineering. 2018.

117. Noble WS. How does multiple testing correction work? Nature biotechnology. 2009;27(12):1135.

118. Perneger TV. What's wrong with Bonferroni adjustments. Bmj. 1998;316(7139):1236-8.

119. Rupert Jr G. Simultaneous statistical inference: Springer Science & Business Media; 2012.

120. Beran R. Balanced simultaneous confidence sets. Journal of the American Statistical Association. 1988;83(403):679-86.

121. Miller R. Simultaneous Statistical Inference. 2 ed: Springer Series in Statistics; 1981.

122. Goodman SN, Berlin JA. The use of predicted confidence intervals when planning experiments and the misuse of power when interpreting results. Annals of internal medicine. 1994;121(3):200-6.

123. Chakrabarti A, Miskovic L, Soh KC, Hatzimanikatis V. Towards kinetic modeling of genome-scale metabolic networks without sacrificing stoichiometric, thermodynamic and physiological constraints. Biotechnol J. 2013;8:1043–57. doi: 10.1002/biot.201300091.

124. Wang L, Hatzimanikatis V. Metabolic engineering under uncertainty. I: Framework development. Metabolic Engineering. 2006;8:133 - 41. doi: http://dx.doi.org/10.1016/j.ymben.2005.11.003.

125. Raue A, Kreutz C, Maiwald T, Klingmüller U, Timmer J. Addressing parameter identifiability by model-based experimentation. IET systems biology. 2011;5(2):120-30.

126. Kiparissides A, Hatzimanikatis V. Thermodynamics-based Metabolite Sensitivity Analysis in metabolic networks. Metabolic engineering. 2017;39:117-27.

127. Dallavilla T, Abrami L, Sandoz PA, Savoglidis G, Hatzimanikatis V, van der Goot FG. Model-driven understanding of palmitoylation dynamics: regulated acylation of the endoplasmic reticulum chaperone calnexin. PLoS computational biology. 2016;12(2):e1004774.

128. Sobol IM. Global sensitivity indices for nonlinear mathematical models and their Monte Carlo estimates. Mathematics and computers in simulation. 2001;55(1-3):271-80.

129. Sarrazin F, Pianosi F, Wagener T. Global sensitivity analysis of environmental models: convergence and validation. Environmental Modelling & Software. 2016;79:135-52.

130. Hsieh N-h, Reisfeld B, Bois FY, Chiu WA. Applying a Global Sensitivity Analysis Workflow to Improve the Computational Efficiencies in Physiologically-Based Pharmacokinetic Modeling. Frontiers in Pharmacology. 2018;9.

131. Verdugo A, Vinod P, Tyson JJ, Novak B. Molecular mechanisms creating bistable switches at cell cycle transitions. Open biology. 2013;3(3):120179.

132. Srinivasan S, Cluett WR, Mahadevan R. Model-based design of bistable cell factories for metabolic engineering. Bioinformatics. 2017;34(8):1363-71.

133. Vital-Lopez FG, Maranas CD, Armaou A, editors. Bifurcation analysis of the metabolism of E. coli at optimal enzyme levels. American Control Conference, 2006; 2006: IEEE.

134. FELL DA, SAURO HM. Metabolic control and its analysis: additional relationships between elasticities and control coefficients. European Journal of Biochemistry. 1985;148(3):555-61.

135. Kiparissides A, Kucherenko SS, Mantalaris A, Pistikopoulos EN. Global Sensitivity Analysis Challenges in Biological Systems Modeling. Industrial & Engineering Chemistry Research. 2009;48(15):7168-80. doi: 10.1021/ie900139x.

136. Bennett BD, Kimball EH, Gao M, Osterhout R, Van Dien SJ, Rabinowitz JD. Absolute metabolite concentrations and implied enzyme active site occupancy in Escherichia coli. Nature chemical biology. 2009;5(8):593.

137. Chubukov V, Gerosa L, Kochanowski K, Sauer U. Coordination of microbial metabolism. Nature Reviews Microbiology. 2014;12(5):327-40.

138. Link H, Fuhrer T, Gerosa L, Zamboni N, Sauer U. Real-time metabolome profiling of the metabolic switch between starvation and growth. Nature Methods. 2015;12(11):1091.

139. Davidi D, Milo R. Lessons on enzyme kinetics from quantitative proteomics. Current opinion in biotechnology. 2017;46:81-9.

140. Davidi D, Noor E, Liebermeister W, Bar-Even A, Flamholz A, Tummler K, et al. Global characterization of in vivo enzyme catalytic rates and their correspondence to in vitro kcat measurements. Proceedings of the National Academy of Sciences. 2016;113(12):3401-6.

141. Cannon WR, Zucker JD, Baxter DJ, Kumar N, Baker SE, Hurley JM, et al. Prediction of Metabolite Concentrations, Rate Constants and Post-Translational Regulation Using

Maximum Entropy-Based Simulations with Application to Central Metabolism of Neurospora crassa. Processes. 2018;6(6):63.

142. Oyetunde T, Bao FS, Chen J-W, Martin HG, Tang YJ. Leveraging knowledge engineering and machine learning for microbial bio-manufacturing. Biotechnology advances. 2018.

143. Schellenberger J, Palsson BØ. Use of Randomized Sampling for Analysis of Metabolic Networks. Journal of Biological Chemistry. 2009;284(9):5457-61. doi: 10.1074/jbc.R800048200.

144. Schork NJ. Personalized medicine: time for one-person trials. Nature. 2015;520(7549):609-11.

145. Nielsen J. Systems biology of metabolism: a driver for developing personalized and precision medicine. Cell metabolism. 2017;25(3):572-9.

146. Geng J, Nielsen J. In silico analysis of human metabolism: Reconstruction, contextualization and application of genome-scale models. Current Opinion in Systems Biology. 2017;2:29-38.

147. Duarte NC, Becker SA, Jamshidi N, Thiele I, Mo ML, Vo TD, et al. Global reconstruction of the human metabolic network based on genomic and bibliomic data. Proceedings of the National Academy of Sciences. 2007;104(6):1777-82.

148. Mardinoglu A, Gatto F, Nielsen J. Genome-scale modeling of human metabolism–a systems biology approach. Biotechnol J. 2013;8(9):985-96.

149. Thiele I, Swainston N, Fleming RM, Hoppe A, Sahoo S, Aurich MK, et al. A community-driven global reconstruction of human metabolism. Nature biotechnology. 2013;31(5):419.

150. Bordbar A, McCloskey D, Zielinski DC, Sonnenschein N, Jamshidi N, Palsson BO. Personalized whole-cell kinetic models of metabolism for discovery in genomics and pharmacodynamics. Cell systems. 2015;1(4):283-92.

151. Long T, Hicks M, Yu H-C, Biggs WH, Kirkness EF, Menni C, et al. Whole-genome sequencing identifies common-to-rare variants associated with human blood metabolites. Nature genetics. 2017;49(4):568.

152. Nicholson JK, Holmes E, Wilson ID. Gut microorganisms, mammalian metabolism and personalized health care. Nature Reviews Microbiology. 2005;3(5):431.

153. Cani PD. Human gut microbiome: hopes, threats and promises. Gut. 2018:gutjnl-2018-316723.

154. Stephanopoulos G. Metabolic fluxes and metabolic engineering. Metabolic engineering. 1999;1(1):1-11.

155. Hadadi N, Hafner J, Shajkofci A, Zisaki A, Hatzimanikatis V. ATLAS of biochemistry: a repository of all possible biochemical reactions for synthetic biology and metabolic engineering studies. ACS synthetic biology. 2016;5(10):1155-66.

156. Hatzimanikatis V, Li C, Ionita JA, Henry CS, Jankowski MD, Broadbelt LJ. Exploring the diversity of complex metabolic networks. Bioinformatics. 2005;21(8):1603-9.

157. Tokic M, Hadadi N, Ataman M, Neves D, Ebert BE, Blank LM, et al. Discovery and evaluation of biosynthetic pathways for the production of five methyl ethyl ketone precursors. ACS synthetic biology. 2018;7(8):1858-73.

158. Nielsen J, Keasling JD. Engineering cellular metabolism. Cell. 2016;164(6):1185-97.

159. Hong K-K, Nielsen J. Metabolic engineering of Saccharomyces cerevisiae: a key cell factory platform for future biorefineries. Cellular and Molecular Life Sciences. 2012;69(16):2671-90.

160. Ferrer-Miralles N, Domingo-Espín J, Corchero JL, Vázquez E, Villaverde A. Microbial factories for recombinant pharmaceuticals. Microbial cell factories. 2009;8(1):17.

161. Schellenberger J, Park JO, Conrad TM, Palsson BØ. BiGG: a Biochemical Genetic and Genomic knowledgebase of large scale metabolic reconstructions. BMC bioinformatics. 2010;11(1):213.

162. Lee JM, Gianchandani EP, Eddy JA, Papin JA. Dynamic analysis of integrated signaling, metabolic, and regulatory networks. PLoS computational biology. 2008;4(5):e1000086.

163. Gonçalves E, Bucher J, Ryll A, Niklas J, Mauch K, Klamt S, et al. Bridging the layers: towards integration of signal transduction, regulation and metabolism into mathematical models. Molecular BioSystems. 2013;9(7):1576-83.

164. Imam S, Schäuble S, Brooks AN, Baliga NS, Price ND. Data-driven integration of genome-scale regulatory and metabolic network models. Frontiers in microbiology. 2015;6:409.

165. Ward PS, Thompson CB. Signaling in control of cell growth and metabolism. Cold Spring Harbor perspectives in biology. 2012:a006783.

166. Ryll A, Bucher J, Bonin A, Bongard S, Gonçalves E, Saez-Rodriguez J, et al. A model integration approach linking signalling and gene-regulatory logic with kinetic metabolic models. Biosystems. 2014;124:26-38.

167. Wittmann DM, Krumsiek J, Saez-Rodriguez J, Lauffenburger DA, Klamt S, Theis FJ. Transforming Boolean models to continuous models: methodology and application to T-cell receptor signaling. BMC systems biology. 2009;3(1):98.

168. Saltiel AR, Kahn CR. Insulin signalling and the regulation of glucose and lipid metabolism. Nature. 2001;414(6865):799.

169. Magkos F, Wang X, Mittendorfer B. Metabolic actions of insulin in men and women. Nutrition. 2010;26(7-8):686-93.

170. Milias-Argeitis A, Rullan M, Aoki SK, Buchmann P, Khammash M. Automated optogenetic feedback control for precise and robust regulation of gene expression and cell growth. Nature communications. 2016;7:12546.

Tuure Hameri

"Looking to apply and further to develop my technical and analytical skills to tackle challenging problems related to different business environments."

Nationality: Swiss, Finnish
Date of birth: 30/09/1990
Telephone: +4122 776 7018
E-mail: tuureh@gmail.com
Address: 8 Ch. Oche-Combe, Founex, 1297, Switzerland

Work Experience

EPFL (Ecole polytechnique fédéral de Lausanne), Lausanne, Switzerland August 2014 – Present
PhD and post-doc at Laboratory of Computational Systems Biotechnology
Research assistant

- Deterministic modelling of the metabolism of bacteria for understanding cellular dynamics
- Developed a workflow for constructing context-specific kinetic models of living organisms
- Devised a procedure that integrates experimental data into models and proposes alternative metabolic engineering decisions, whilst accounting for uncertainty
- Implemented a sensitivity analysis method that facilitates design of experiments and allows the reduction of uncertainty/risk in kinetic models
- Initiated interdisciplinary collaboration with Université de Lausanne for applying statistical methods for ranking and comparing candidate metabolic engineering decisions
- Presented research in international conferences

Teaching assistant

- Courses taught: introduction to chemical engineering, programming in MATLAB, advanced principles and applications of systems biology
- Delivered lectures and prepared/corrected examinations in French and English

Invensys (FTSE 100), Crawley, UK June 2012 – August 2012
Global technology company providing process automation, control and monitoring solutions; *Sales Intern*

- Validated change management utility for engineering tools
- Produced training collateral for delivery to enable incremental revenue
- Designed branding templates that were adopted by the corporate marketing team
- Drew technical network system architecture diagrams used by sales department in proposal quotes

SABIC – IP, Bergen Op Zoom, Netherlands June 2011 – July 2011
Saudi Basic Industries Corporation Innovation Plastics ($40 billion revenue) is an international company producing chemicals, metals, fertilizers and plastics; *Chemical Operations Intern*

- Interviewed around 30 employees from Chemical Operations department to make an inventory of the needs for advanced training and simulations to allow migration to a new process control system (DELTA–V)
- Collaborated with the Process Automation department to determine what training and simulation tools are already available and estimated the cost of the identified further requirements
- Assisted operator shifts to gain understanding of process control
- Delivered a report and presentation on the findings for the plant management team

Addax Petroleum, Geneva, Switzerland June 2010 – September 2010
International oil and gas exploration and production firm ($5 billion revenue); *Facilities Intern*

- Simulated and optimized oil separation processes to minimize hydrate formation using Aspen HYSYS software
- Calculated pressure drops and flow velocities in pipes to ensure compliance with European safety standards
- Tracked variations in reservoir gas lift to determine if changes in water injection were required
- Participated in weekly management meetings to report results

Atacama Labs, Geneva, Switzerland July 2009 – August 2009
Pharmaceutical start up company specialized in dry granulation processes; *Intern*

- Researched patents with similarities to the dry granulation process in question
- Filed a report on patents that could possibly cause copyright and technical infringement issues

Education

EPFL (Ecole polytechnique fédéral de Lausanne), Switzerland August 2014 –December 2018

- PhD. in Chemical Engineering
- Thesis: Towards comprehensive and consistent kinetic models of metabolism under uncertainty

University of Cambridge, UK September 2012 – August 2013

- M.Phil. in Advanced Chemical Engineering, 76% Average
- Courses: Computational Fluid Dynamics, Numerical Methods, Fluid Mechanics and Project Management
- Thesis: Optimal Nozzle Geometries for Zero Discharge Fluid Dynamic Gauging, Distinction
- Group consultancy project for GlaxoSmithKline, a global healthcare business
- Management of Technology and Innovation certificate

University College London, UK September 2009 – June 2012

- B.Eng. in Chemical Engineering, First Class Honours
- Courses: Mathematics for Engineers, Computer Aided Design and Modelling, Process Engineering
- Thesis: Biodiesel plant group design project

International School of Geneva, Switzerland September 1997 – June 2009

- Bilingual International Baccalaureate (39/45 Points)
- IGCSE (Mathematics A* and Geography B)
- Awarded the Honour Role in the 3 final years for maintaining high academic performance

Additional Skills

Languages: Finnish (Native), English (Fluent), French (Fluent) and German (3 years of study)
Computer Skills: Microsoft Office, MATLAB, Polymath, GAMS, Aspen HYSYS, COMSOL and ChemCAD
Hobbies: Travelling, CrossFit and rowing (Qualified for county–level championships)

References Upon Request

Ce document a été imprimé au Centre d’impression EPFL,
imprimerie climatiquement neutre, certifiée myClimate.