

New and improved methods for integrated models of metabolism and gene expression

Présentée le 12 février 2021

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

Pierre Guy Rémy SALVY

Acceptée sur proposition du jury

Prof. A.-C. Corminboeuf, présidente du jury
Prof. V. Hatzimanikatis, directeur de thèse
Dr A. Singh, rapporteuse
Prof. S. Finley, rapporteuse
Prof. D. Kuhn, rapporteur

Everybody has a capacity for a happy life.
All these talks about how difficult times we live in,
that's just a clever way to justify fear and laziness.
— Lev Landau

Je ne sais pas ce qui est beau, mais je sais ce que j'aime
et je trouve ça amplement suffisant.
— Boris Vian

*Pour Simone, ma grand-mère qui m'a toujours dit que même si j'avais des bonnes notes
à l'école, je n'irais pas bien loin sans un peu de bon sens.*

—

*For Simone, my grandmother who always told me that even if I had good grades at school,
I would not go very far without a bit of common sense.*

Acknowledgements

This thesis would not have been what it is, had I not been surrounded by beautiful, insightful, talented, and exceptional people. Here is a small tribute to them.

First and foremost comes Prof Vassily Hatzimanikatis, a.k.a. *the Boss*. Vassily, thank you for your academic and non-academic guidance, your countless stories and anecdotes that spiced our days at the lab. You also managed, somehow, to provide me with good oversight while giving me plenty of freedom in my research¹. I also appreciate the ability you taught me, to shed the complicated maths lingo and relate better to the biology behind it all. This differentiating skill between a hands-on computational biologist and an exiled esoteric hermit, I owe it to you. And thank you for the numerous quality debates around equally good tea, from theology and practical philosophy to computational topics, and the many mental exercises in relating the rationalist and existentialist, Catholic and Orthodox, Stoic and Epicurean, cataphatic and apophatic, schools of thought — among many others — in contexts both ancient and modern. Also I think we should both thank Boris Vian.

I would like to express my gratitude towards my thesis committee, to whom I had the honour to present my work. Dr Amoolya Singh, you have been as much a mentor as a role model since I had the chance to meet and work with you in those sunny days in California. Prof. Daniel Kuhn, my *de facto* mentor at EPFL, thank you for your ever insightful advice, in discussing thesis orientation in the early days, or in the course I took with you, and for your quick-witted humor in more informal settings. Prof. Stacey D. Finley, thank you for your sharp questions and thoroughness in examining my thesis, but above all thank you for showing me how people of science can also be actors in societal issues. Finally, I would like to thank you, Prof. Clémence Corminboeuf, for being such a deft, agile, determined and confident Master of Ceremony in this logistically complicated thesis defense.

I would never have been admitted in Vassily's lab without the right people to encourage

¹To do justice to History, I have to admit that when Vassily asked me how I felt about working on ME-models, after 1 year of PhD, I more or less turned off the offer. I said there was not much left to do and they were too complicated.

Well, here we are now, and the next 100+ pages are talking about it.

Acknowledgements

and support me. Many heartfelt thanks in this regard to Joshua Lerman, Olivier Rolland, Amoolya Singh, Vineet Rajgarhia. You all participated in stirring my way towards this PhD, and not only provided me with an infinite amount of wise advice, but also made me ask the right questions to make my choice. And huge thanks to Véronique Stoven for being the one person who opened the gates of Biotech to me.

Dr. Ljubiša Mišković, or *Misko* – you were the Serbian friend and sage and scientist and party buddy I did not know I needed. Хвала лепо for all these good moments, teaching reluctant students the wonders of automated control, sampling beers at Sat, eating crab in Seattle, rating pastries in Braga; but also for all the thoughtful conversations, the constant perfect advice, your daily joviality and good humor, and the countless “in Serbia we have this joke ...”. My PhD would have definitely been less colorful without you.

I also had my lab crew, my partners in crime in all things lab-related. Maria, you’ve been both the fiery fires of determination and the ice buckets of realism that accompanied me throughout my journey, and helped me achieve my destination without straying from my path. Thank you for being my no-questions-asked accomplice, and also my no-nonsense ward. Daniel, I’ll use your words: thank you, too, for having that twisted, non-differentiable humor. I felt much less alone with you as a sparring buddy for convoluted thought experiments, and I felt endorsed when you’d follow me in my most absurd endeavours. Robin, thank you for chanting to me the gospels of the Open Source, and the endless discussions on social, technical, and socio-technical topics. And thank you for the numerous unnameable breakfasts. Jasmin, you are the chilliest of all, and you are also my last mate from the olden times. Thank you for your contagious peaceful take on things, and of course for the great climbing lessons.

The rest of LCSB also deserves a big thanks for being, all those years, a welcoming family and a cozy place to work at. The old-timers, Milenko, Tuure, Anush, Meriç, Tiziano, Yves, who were my first contact in the lab, and whom I was happy to meet again. Georgios Fengos, man, what luck I had to work with you, and to get to know you! Thank you for everything you know. Big thanks to the salad club, for all these yummy lunches, which definitely contributed to my (reduced) efficiency in the afternoons. From my era, Sofia, you were a fun coworker to have, and thank you for the many songs we shared, and your quality tour-guiding in Greece. Homa, we started the same day, and we flew so many times to Denmark to meet these PAcMEN people, thank you for being the joyful travel companion and workmate on these multiple occasions. Kaycie, thank you for all the time you spent reading my no-so-good manuscripts, your awesome, straight-to-the-point briefs, and your eagerness to party. Then comes the trio of the new kids club, Anastasia (thank you for the Russian lessons!), Asli (you’re the new Pope of Code now), and Evangelia (thank you for being the party catalyst of the new generation). With you three, I have not doubts that the social life of the lab will be in good hands. Also big thanks to the rest of the team, Zhaleh, Liliana, and Omid.

Thanks a lot to the H2020 PAcMEN team, and in particular to Anna, Mu-En, Roy, Hélén, Vasil, Paul, for the fun moments in Denmark, Switzerland, Germany and Portugal.

Après viennent les athlètes indépendants.

Olivier, t'es le mec qui ne s'arrête pas, et tu m'as beaucoup inspiré sur comment gérer l'écriture de la thèse quand, lors du même déjeuner, tu m'as annoncé en série avoir commencé à rédiger, t'être inscrit à un triathlon pour lequel tu n'avais pas commencé à t'entraîner, et avoir commencé à prendre des cours d'allemand intensifs. Je ne te cache pas que la barre était haute pour tenter d'être aussi *cool* que toi. Merci pour les millions de cafés et (petits-)déjeuners qui ont rythmé nos vies académiques et professionnelles après qu'on se soit rencontrés au ski-maths de Zinal.

Speaking about Zinal, I met in this place some magical people I would have had no chance to meet if we had not shared a fondness for optimization and skiing. Christoph, Çağil (Charlotte), Kilian, thank you for sharing these precious moments in *The Bar* and *The other Bar*, at Sat, or by the lake. Our endless train of complaints, and moments of joy and glory when papers came out, hardship and bliss on the snow, are memories I will hold dear forever. Kilian, lots has been left unsaid between us. You are my absolute role model in so many aspects it is hard to make an exhaustive list. I am grateful I got to know you in my life, and I hope we meet in the next. We still need to formulate this robust metabolism problem.

Une chose importante lorsque l'on travaille de façon obsessive-compulsive sur un sujet, c'est de vivre avec des gens capables de nous sortir la tête du guidon quand il le faut. En particulier, en vivant en colocation, j'ai eu la chance de rencontrer tellement de gens différents qui ont su me faire changer d'air. David, t'es le premier ami suisse que j'ai eu. Merci de m'avoir fait découvrir tant de choses quand je n'étais encore qu'un frouze parisien fraîchement débarqué du train. Et surtout merci de m'avoir fait confiance en reprenant ta coloc. Tao, Matti, Cyprien, Guillaume, merci d'avoir contribué à la grandeur de Château Cuiller, la meilleure coloc du 1022.

Paul, on n'a jamais vraiment été colocs mais on a partagé plein d'autres choses. Merci d'avoir eu l'idée d'organiser ces soirées dégustation de vins et liqueurs de qualité, mais surtout merci infiniment de m'avoir ouvert la porte de la coloc du 33, où j'ai passé une incroyable dernière année de thèse. Manue, merci pour tous ces moments de chill, la décomplexation des journées lecture/tisane enveloppées dans un plaid au fond du canapé; et aussi, merci pour ces trucs absurdes comme aller acheter des vêtements vintage au kiloshop. Cléa, merci pour la diversité de ta conversation, la richesse de ta culture, les innombrables moments où tu as su me faire considérer avec un oeil différent des sujets sociétaux pendant nos débats; mais aussi et surtout merci de ta rafaraîchissante bienveillance naturelle. Pierre (le Jeune), merci d'être un coloc aussi hipster inavoué que moi, merci de m'avoir embarqué dans la grimpe en falaise, merci d'avoir instauré la

Acknowledgements

tradition des gauffres en soirée, merci de m'avoir chauffé pour faire de la bière avec toi; bref, merci du grin de folie et de chaos que tu as pu m'apporter. Roeltje, sis, you were my favorite discovery of the confinement. Building a swing, planting potatoes, gurken, tomatoes, peppers, playing water-polo together, getting me to hike to camp and climb a bunch of via ferratas, dive-fishing for lost keys in the lake, those are just some of the very many weird or cool (or both) things I never thought I'd do but I am happy we did together. Thank you so much for being my best (Dutch) friend, for the unchecked night discussions, and for being my partner in crime in Divinity.

Cette coloc était probablement un des meilleurs endroits où attendre que ne passe la tempête qu'était l'année 2020.

N'importe quelle personne ayant passé plus de 30 secondes à me parler a entendu parler d'eux, les zigotos du water-polo. Merci infiniment à ce merveilleux groupe de personnes qui m'a procuré, paradoxalement, de grandes bouffées d'oxygène hors de l'EPFL. Noémie et Pauline, mes zouzes de la 3, vous avez été mes soleils constants dans l'eau comme en dehors. Merci pour tous les moments incroyables qu'on a passés ensemble, au ski, dans les montagnes, en match sous la pluie, autour d'une fondue, sur des chaises longues, devant des films, et surtout au non-Cambodge. No, merci pour ton écoute infinie et ton intelligence émotionnelle hors normes. Pau, merci pour les milliards de discussions où souvent tu as su me faire penser différemment.

Borja, t'es le bro que j'ai jamais eu, le roi des plans absurdement géniaux, et l'incarnation de la génialité spontanée. Merci pour les séances de polo du midi dans le lac, les team-plays et pro moves en match, ta bonne humeur infinie, la cueillette aux champignons, la peinture, et le reste.

Merci à la meilleure team de la ligue régionale, si pas en score, au moins en ambiance: Jean, Benoit, David, Justine, Christoph, Fanny, Thomas, Filip, Agron, Camilo, et bien sûr Wouter. Surtout, merci infiniment à Seb de nous supporter depuis toujours, malgré tout. Mention spéciale pour Davor, coach remplaçant à ses heures perdues, mais surtout meilleur ski-guide de Suisse et formidable amphitryon à maintes reprises.

Les Tsunas, vous êtes ma seconde famille, et vous avez été là à chaque instant de ma thèse grâce à la magie de la messagerie instantanée. Merci d'être cette même bande de trublions protéiforme, apte à décortiquer des sujets techniques comme sociétaux en moins de temps qu'il ne faut pour prononcer vos tsunanymes, mais aussi d'être ce formidable bouillon d'humour et d'amour sans limite tangible. Afin de préserver votre anonymité, j'utiliserai seulement vos surnoms du moment, tsunanymes éphémères et toutefois héritiers de plus de 5 ans de tradition. Merci à Detective Hunter C15 SuprêmeLeaderGrilleDawg; VP du monde Girlfrisch Cachotier JeanneAuSecours; Dominator Hubert Lilou, coupdegenix en burkini; La courbature, dormeur sur chien Clavette de champ; Poil de sillon mam, bayvetronic de quiche; Dragon Pual, discoStool & loukoum; Patient 1 debite, PONÇAGE

die dalai-kebab; Tonton Tipi Prouxballe, Ministre, Petite Tornade; la Poule quantique chicken fight; Joseph Nutscaper, tôt la idiot, the Coqjammer. Mention spéciale pour B. Chabot, aka Binistre d'abbelbakayan, 666 le diable, notre fils spirituel à tous, qui a su à maintes reprises nous faire profiter de son sens de l'humour étrangement alien et familier à la fois.

Tout ce travail aurait une autre allure sans Swaglang, mon crew et think-tank depuis dix ans maintenant. La combinaison d'activités non conventionnelles et de discussions bien techniques a provoqué la germination de bien des idées, souvent importantes, utilisées dans ma recherche. Léo, merci d'avoir été mon PhD-tes-not-rants buddy toutes ces années, mon expert de confiance en IA, mon bon conseil technique, mon bro de Californie, et surtout l'hôte de nos nombreuses semaine de travail dématérialisé à St Malo. Freffy, merci d'avoir été le Jack-of-all-trades et Master-of-all, source de savoir inépuisable sur tant de sujets, qui nous a appris à brûler du bois selon une technique traditionnelle japonaise, mais aussi le cocktail-bar master de Paris, et l'hôte de chaque Nouvel An. Clément, merci d'avoir été mon gourou software engineering à distance, beaucoup de tes anecdotes, conseils ou idées ont fini d'une façon ou une autre dans mon code. Dac, merci d'avoir été un paragon du chill, et un co-conspirateur du kouign-amann de l'hérésie. Guillaume, merci d'avoir été mon point de contact à Boston, et de m'avoir couvert pendant mon infiltration en tant qu'agent russe dans le milieu brésilien. Victoire, merci d'être une meuf si unique dans ton humour, si enjouée, ma partenaire de rock préférée, et une oreille si attentive. Et bien sûr, merci aussi à Camille, Michel & Hélène, d'être juste de si chouettes amis.

Lise & Laura, mes invitées de qualité, c'était toujours un plaisir de vous avoir à la maison. Et Lise, merci pour tous ces bons restaurants où l'on a pu tant échanger sur la vie lors de mes séjours à Paris.

David, merci d'avoir été mon plus vieux pote depuis 20 ans maintenant. Clémence, tu n'es pas loin derrière. Merci à vous deux pour les bons moments passés à Nantes.

JB, merci d'avoir été là si souvent pour prendre des nouvelles, t'assurer que je ne sombrais pas dans la folie, d'avoir aidé à la conception de notre première bière la SalChevreMol, et d'avoir été au rendez-vous aux quatre coins du globe lorsque le hasard du business nous faisait atterrir au même endroit.

Christian, merci d'avoir été un si bon conseil sur tant de choses et depuis si longtemps, et merci de ta culture et curiosité scientifique, qui ont initié bien des discussions dans ton atelier. Chacune des nos créations est empreinte de nos divagations sur le monde, et me rappelle ces moments de qualité partagés avec toi.

Margaux, tu as été un monument central de ces années en Suisse. C'est vrai que c'était un peu différent de la Californie. Merci pour toutes ces aventures qu'on a partagées, merci de m'avoir motivé à faire de la photo, merci pour ces millions de gâteaux tellement

Acknowledgements

bons que je ne veux plus aller dans une pâtisserie, merci de m'avoir appris à faire du ski, merci de m'avoir écouté quand j'en avais marre, merci d'avoir célébré avec moi toutes ces petites victoires dans la vie de doctorat, merci de m'avoir fait tant rire. Bref, merci d'avoir été une coéquipière si formidable pendant ces quatre ans.

Et le plus important pour la fin, je voulais remercier mes parents de s'être battus pour que j'aie l'opportunité de suivre mes intérêts personnels et scientifiques, tout en me laissant la liberté de faire ce qui me plaisait, sans jamais questionner mes choix. Merci à ma soeur Anne, qui a toujours été ma plus fervente supportrice. Merci Simone de m'avoir montré le sens d'avoir du bon sens. Et la recette des crêpes.

And my thanks to you, reader, for having mustered the courage to read some of my work!

Lausanne, September 21, 2020

P. S.



Abstract

The beginning of the 21st century was marked by the advent of disruptive technologies, which ushered an era of groundbreaking advances in fundamental sciences, carried by the great pace at which computational capabilities spread and evolved. But the new century also came with its fair share of challenges. Anthropogenic climate change brought issues of sustainability in the production chemicals and food. The global improvement of life expectancy and diagnostic methods also saw the increased incidence of illnesses for which age is a risk factor, such as cancer and dementia.

These challenges share a connection to living matter, and understanding and improving biology are two goals of systems biology and metabolic engineering. These fields provide tools and methods that are suited to respond to the new requirements of chemical production, food availability, and health and medicine through the understanding and engineering of living cells. Engineered microorganisms are already used in the production of both commodity and specialty chemicals, genetically improved crops are a possible answer to the ever-increasing food demand, and new medical treatments rely on an improved understanding and control of cellular idiosyncrasies.

Efficient engineering requires mathematical models. Over the last decades, the increasing availability of full genome sequences and their translation into models of metabolism enabled the emergence of a wide gamut of methods to describe the inner workings of the cells we study. In particular, models of metabolism and gene expression (ME-models) were the first formulation to account simultaneously for cell metabolism, and the expression mechanisms translating genetic information into proteins.

In this thesis, I present a new, and improved, formulation for ME-models, and apply it to elucidate the emergence of non-trivial elements of cell physiology. This new ME-model formulation, ETFL, allows the integration of more experimental data than the previous state of the art, while being more efficient than previously published equivalent methods. Then, I show ETFL elucidates complex cellular behaviors. In particular, I demonstrate the preferred consumption of specific carbohydrate by *E. coli*, or diauxie, is the result of an optimal program of the cell towards growth, under the constraints of proteome limitation. I show that ETFL can be adapted to elucidate the dynamics of the proteome in the

Abstract

cell and the transition from one physiology to another. I also describe the construction of a ME-model of a eukaryotic organism, *S. cerevisiae*, and how the model produced can account for the emergence of overflow metabolism, or the Crabtree effect. Finally, I build a model of human colon cancer, and present a formulation for ETFL that allows to account for regulatory interactions in ME-models. I use the model to reproduce the known mechanisms of action of the drug metformin, and show it has a dual, dose-dependent action. I also show how such models can be used to predict potential mechanisms of resistance against treatment. In a second part of this thesis, I present open source software pieces I developed and contributed to, to promote open science.

This work outlines the potential for ME-models in systems biology, and shows how to use them to elucidate complex cellular physiologies. The methods presented in this work also show how these new and improved ME-models constitute a major step towards systematic, integrated whole-cell modeling.

Keywords

Metabolism, Gene expression, Genome-scale models, ME-models, Data integration, Constraint-based modeling, Dynamic models, Gene regulation, Drug mechanisms

Résumé

Le tournant du 21^{ème} siècle fut marqué par l'avènement de technologies disruptives à l'origine de progrès considérables dans les sciences fondamentales, et ce au rythme de l'expansion et de l'évolution de la puissance de calcul à notre disposition. Mais le nouveau siècle n'est pas arrivé seul : en son sillage, de nouvelles problématiques. Les changements climatiques dus à l'Homme questionnent la durabilité de nos procédés chimiques et agricoles. L'amélioration généralisée de l'espérance de vie et des méthodes de diagnostic sont également la cause d'une augmentation de l'incidence de maladies liées à l'âge, comme différents types de démences et cancers.

Ces problématiques sont liées par leur rapport au vivant. Mieux comprendre la biologie pour l'améliorer, sont deux objectifs de la biologie des systèmes et de l'ingénierie métabolique. Ces domaines de la science nous procurent des outils pour répondre aux nouvelles attentes dans les domaines de la chimie industrielle, l'agriculture, et la santé. L'utilisation d'organismes modifiés fait maintenant partie de nombreux procédés chimiques, plusieurs plants génétiquement améliorés sont en développement pour répondre à la demande croissante en denrées alimentaires, et de nombreuses nouvelles thérapies sont tributaires d'une compréhension et d'un contrôle avancés de cellules spécifiques.

Une bonne ingénierie requiert une bonne modélisation mathématique. Au cours des dernières décennies, le nombre croissant de séquences génétiques et de modèles dérivés à notre disposition ont alimenté un vaste écosystème de méthodes décrivant le fonctionnement intime des cellules vivantes. Parmi celles-ci, les modèles du métabolisme et de l'expression génétique (ME-modèles) furent les premiers à formaliser les liens entre le métabolisme d'une cellule et l'expression de ses gènes.

Dans cette thèse, je présente une formulation nouvelle, améliorée des ME-modèles, et l'utilise afin d'expliquer l'émergence d'éléments non-triviaux de la physiologie cellulaire. Cette nouvelle formulation, ETFL, facilite l'intégration de plus de données expérimentales que le précédent état de l'art, tout en étant plus efficace. Ensuite, je montre que ETFL explique certains comportements cellulaires complexes. En particulier, je montre que la consommation préférentielle de certains sucres, ou diauxie, par *E. coli*, est la conséquence d'un programme optimal de croissance de la cellule sous contrainte d'un protéome limité. Je

Résumé

décris également la construction d'un ME-modèle pour l'organisme eukaryote *S. cerevisiae*, et comment ledit modèle explique l'émergence d'un métabolisme excédentaire, aussi appelé effet Crabtree. Enfin, je construis un modèle de cellule humaine du cancer du côlon, et présente une extension de la formulation de ETFL permettant de modéliser certains effets dus à la régulation cellulaire. A l'aide de ce modèle, je reproduis l'effet de la metformine, médicament préconisé dans le traitement du cancer du côlon, et démontre un effet double suivant son dosage. Je montre également comment utiliser ce type de modèle pour prévoir de potentiels mécanismes de résistance thérapeutique. Dans une seconde partie de ma thèse, je présente plusieurs logiciels que j'ai conçus, ou au développement desquels j'ai participé. Ces logiciels sont libres, et leur code source est accessible à tous, afin de soutenir une science plus accessible.

Ce travail met en exergue le potentiel des ME-modèles dans le domaine de la biologie des systèmes, et explique leur utilisation afin de comprendre certains aspects non triviaux de la physiologie cellulaire. Les méthodes présentées dans cet ouvrage montrent également que cette formulation nouvelle et améliorée des ME-modèles constituent un pas décisif vers la construction de modèles cellulaires plus complets et exhaustifs.

Mots-clefs

Métabolisme, Expression génétique, Modèles à l'échelle du génome, ME-modèles, Intégration de données, Modélisation sous contraintes, Modèles dynamiques, Régulation génétique, Mécanisme d'action

Contents

Acknowledgements	i
Abstract (English/Français)	vii
List of Figures	xvii
List of Tables	xxiii
Introduction	1
An informal prologue	1
Field-specific context and definitions	3
Modeling the cell at the level of metabolism and expression	6
Motivations	9
Structure of the thesis	10
 I Integrating expression mechanisms in genome-scale models	 13
1 The ETFL formulation allows multi-omics integration in thermodynamics-compliant metabolism and expression models	17
1.1 Introduction	19
1.2 Results and Discussion	21
1.2.1 Formulation of the expression problem	21
1.2.2 Application: <i>E. coli</i> genome-scale model iJO1366	27
1.2.3 Performance	37
1.2.4 Adaptation of FBA-based methods to ETFL	39
1.2.5 Building an ETFL ME-model for other organisms	41
1.3 Conclusion	42
1.4 Materials and Methods	43
1.4.1 Preliminaries, Conventions, and Notations	43
1.4.2 General constraints for enzymes	45
1.4.3 Constraints specific to Ribosomes	46
1.4.4 Constraints specific to RNA Polymerase	47
1.4.5 Constraints for Peptides	48
	xi

Contents

1.4.6	Constraints for mRNAs	49
1.4.7	Constraints specific to rRNAs	51
1.4.8	Constraints specific to tRNAs	51
1.4.9	Reformulation of the bilinearity of the problem	52
1.4.10	Approximation of the growth rate	53
1.4.11	Linearizing the bilinearity	55
1.4.12	Petersen linearization	55
1.4.13	Discretization of mRNA and enzyme content	57
1.4.14	Discretization of DNA content	58
1.4.15	Gene copy number and RNAP saturation	60
1.4.16	Expression constraints for genes without enzymes	61
1.4.17	Scaling	61
1.4.18	Advanced modeling	62
1.4.19	Thermodynamics-based constraints	66
1.4.20	Data	68
1.4.21	Model modification	68
1.4.22	Enzyme estimation	69
1.4.23	Essentiality analysis	69
1.4.24	Hardware	70
1.5	Code availability	70
1.6	Data availability	70
2	Dynamic ME-models: Emergence of diauxie as an optimal growth strategy under resource allocation constraints in cellular metabolism	73
2.1	Introduction	75
2.2	Results	77
2.2.1	Conceptual model for the emergence of the diauxie phenotype from proteome limitation	77
2.2.2	Diauxie in genome-scale, ME-models with thermodynamic constraints	81
2.3	Discussion	90
2.4	Material and Methods	92
2.4.1	Rate of change of fluxes	92
2.4.2	Variability in the estimation of macromolecule concentrations . .	93
2.4.3	Backwards Euler integration scheme	94
2.4.4	Chebyshev center	95
2.4.5	Initial conditions	96
2.4.6	Extracellular concentrations	97
2.4.7	Model	97
2.4.8	Kinetic information	97
2.4.9	Implementation	98
2.5	Supporting Information Appendix (SI)	98
3	Emergence of the Crabtree effect in a model of metabolism and ex-	

pression for <i>S. cerevisiae</i>	101
3.1 Introduction	103
3.2 Results and Discussion	104
3.2.1 Model description	104
3.2.2 Prediction of specific growth rate	105
3.2.3 Gene essentiality analysis	107
3.2.4 Crabtree effect	108
3.3 Conclusion	109
3.4 Materials and Methods	111
3.4.1 Formulation of the ETFL model	111
3.4.2 Ribosomes and RNA polymerases	112
3.4.3 Allocation data and constraints	114
3.4.4 Modifying the growth-associated maintenance	116
3.4.5 Gene-protein-reaction coupling	117
3.4.6 Gene essentiality analysis	117
3.4.7 Chemostat simulations	119
3.4.8 Code and Dependencies	119
4 Dose-dependent drug effect and resistance mechanisms in a cancer model of metabolism and expression	121
4.1 Introduction	123
4.2 Results and Discussion	125
4.2.1 Experimental setup	125
4.2.2 Experimental results	125
4.3 Conclusion	131
4.4 Materials and Methods	132
4.4.1 Condition-specific cancer ME-models	132
4.4.2 ETFL	133
4.4.3 Gene copy number and RNAP saturation	134
4.4.4 Incorporation of signaling pathways	134
4.4.5 Representative solution	138
4.4.6 Optimization procedure	140
4.4.7 Parameters	141
4.4.8 Data and code availability	141
II Open Science efforts	143
5 pyTFA and matTFA: A Python package and a Matlab toolbox for Thermodynamics-based Flux Analysis	147
5.1 Introduction	150
5.2 Materials and methods	151
5.2.1 Embedding thermodynamic constraints	151

Contents

5.2.2	Implementation	152
5.3	Usage	152
5.4	Conclusion	153
6	A Python implementation of the metabolic network analysis and reduction algorithms redGEM and lumpGEM	155
6.1	Introduction	157
6.2	Materials and methods	158
6.2.1	Reduction and lumping	158
6.2.2	Thermodynamics	159
6.2.3	Implementation	159
6.3	Usage	159
6.4	Conclusion	160
7	Symbolic Kinetic Models in Python: SKiMPy	161
7.1	Introduction	163
7.2	Material and Methods	163
7.2.1	Symbolic kinetic models	163
7.2.2	Sampling steady state consistent parameter sets	164
7.2.3	Serialization	165
7.2.4	Implementation	165
7.3	Usage	165
7.4	Conclusion	166
	Concluding remarks	169
	Conclusions	169
	Outlook	171
III	Appendix	175
A	ETFL: Supplementary information	177
B	dETFL: Supplementary information	205
C	Regulation-enabled ETFL models of cancer: Supplementary information	219
D	pyTFA & matTFA: Supplementary information	223
E	SKiMPy: Supplementary information	229
	Bibliography	252

Glossary	253
Curriculum Vitae	257

List of Figures

- 1.1 Growth rate with respect to glucose uptake for differently constrained models in the ETFL framework. Legend in the same order as the height of the right-most point of each curve in each figure. **a.** Growth rate predictions using the EFL, ETFL, vEFL, vETFL models (dark blue, light blue, purple, orange); **b.** Growth rate predictions accounting for missing enzymes using vETFL (orange) and models (i)-(iii) (purple, light blue, dark blue) representing different initial enzyme assumptions, with k_{cat} values obtained from vETFL or $k_{\text{cat}} = 172 \text{ s}^{-1}$, and with/without inferred enzymes. Lines have transparency to better see overlaps. 31

- 1.2 Concentration variability of peptide species, sorted by average peptide concentration (darker disc). Lower bounds that were 0 were set to the accuracy of the solver, 10^{-9} . The horizontal line on the left side of the figure represents ribosomal peptides, which is narrow due to their instrumental role in making the tightly constrained amount of protein in the cell at a given growth rate. The vertical line in the middle represents the dummy peptide, which accounts for unmodeled peptides (non-metabolic proteins and enzymes with missing information) and therefore is used by the solver as a slack. 32

- 1.3 Voronoi map of the predicted abundances of mRNAs. Each colored patch represents a different mRNA, with its area in proportion to its relative abundance. Genes can be clustered using KEGG Gene Ontology (GO) terms. Colors indicate the clustering. 34

1.4	Confusion matrices for gene essentiality studies. a. Conventions from Orth <i>et al.</i> (48) for gene essentiality. TN is True Negative. FN is False Negative. FP is False Positive. TP is True Positive. The color shading represents how good the classification is in the experimental class. Perfect classification should have a strict red first diagonal, as shown on this example. b. Gene essentiality prediction for the FBA model iJO1366, yielding a Matthew's correlation coefficient (MCC) of 0.69. c. Gene essentiality prediction for the genes expressed in the vETFL model, yielding a MCC of 0.65. d. Gene essentiality prediction for the vETFL model, where genes without enzyme assignment were tested using gene to protein to reaction (GPR) associations from the iJO1366 model, yielding a MCC of 0.61. e. Gene essentiality prediction for the vETFL model with estimated enzymes with all $k_{\text{cat}} = 172 \text{ h}^{-1}$, yielding a MCC of 0.54.	35
1.5	Histograms displaying the distribution of solving times of each type of model during the data generation for this study. The darker area represents data between the 5th and 95th percentiles.	38
1.6	Discretization example for specific growth rate and growth-dependent parameters. a. Discretization of μ into $\hat{\mu}$. The step approximation transforms the continuous interval $[0, 2.5]$ into the discrete set $\{0, 0.25, \dots, 2.5\}$. b. Example of piecewise linear interpolation and discretization of the protein mass ratio from Neidhardt <i>et al.</i> (47). Red circles represent the values reported. The dashed line is the piecewise linear interpolation. The solid line is its discretization.	59
2.1	a. Conceptual model used for the preliminary analysis, where "glc" stands for glucose, "lcts" for lactose. The catalytic efficiency of the enzymes are assumed to be the same. Three enzymes are assumed to be necessary to produce the intermediate metabolite G6P from lactose, and only one enzyme is required from glucose. b. Optimization problem used to represent the model. v are fluxes, E are enzyme concentrations, E^0 are reference values, MW are molecular weights, ρ is the mass fraction of the cell occupied by the enzymes we consider, \dot{E}_{max} is the maximal variation of enzyme concentration over time, and dt is the integration interval. c. Enzyme content over time for the conceptual model growing on a mixed substrates. d. Changes in sugar content of the batch reactor over time.	79

2.2	Comparison of simulated and experimental data of glucose depletion over time (Varma and Enjalbert). Simulated data represented by a solid line, experimental data by crosses. a. Temporal evolution of the simulated extracellular concentrations of glucose, and acetate (full lines), versus experimental data (Varma dataset) (crosses). b. Temporal evolution of the simulated (full lines) extracellular concentrations of glucose, and acetate (solid lines), versus experimental data (Enjalbert dataset)(crosses). c. Cell concentration (full line) and growth rate (dashed line) over time, simulation and Varma dataset (crosses). d. Cell concentration (solid line) and growth rate (dashed line) over time, simulation and Enjalbert dataset (crosses). * Experimental values were in optical density (OD600), and were linearly scaled to represent cell concentrations.	84
2.3	Possible uptake routes for glucose (blue) and lactose (orange), towards glucose-6-phosphate (G6P). The splitting of lactose by LACZ can be done either intracellularly, or in the periplasm. Routes towards the main central carbon metabolism are in gray. Figure made using Escher. (91)	86
2.4	Diauxic simulation with glucose-only preculture: a. Temporal evolution of the extracellular concentrations of glucose (blue), lactose (orange), and acetate (green). b. Exchange rates of the cell. Positive exchange rates mean production, negative exchange rates mean consumption. c. Cell concentration (solid line) and growth rate (dashed line) of the culture over time. d. Mass of enzymes allocated to the transformation of glucose (blue) and lactose (orange) in G6P. The dashed gray line shows the levels of β -galactosidase (LACZ) enzyme (in the lactose pathway).	87
2.5	Diauxic simulation with lactose-only preculture: a. Temporal evolution of the extracellular concentrations of glucose (blue), lactose (orange), and acetate (green). b. Exchange rates of the cell. Positive exchange rates mean production, negative exchange rates mean consumption. c. Cell concentration (solid line) and growth rate (dashed line) of the culture over time. d. Mass of enzymes allocated to the transformation of glucose (blue) and lactose (orange) in G6P. The dashed gray line shows the levels of β -galactosidase (LACZ) enzyme (in the lactose pathway).	89
3.1	Growth rate of the yETFL models (orange and blue) and Yeast8 FBA (dashes) with respect to glucose uptake. Experimental data (crosses) by Van Hoek <i>et al.</i> (135)	107

List of Figures

3.2	Confusion matrices for gene essentiality studies. a. Conventions from used for gene essentiality. TN is True Negative. FN is False Negative. FP is False Positive. TP is True Positive. The color shading represents how good the classification is in the experimental class. Perfect classification should have a strict red first diagonal, as shown on this example. b. Gene essentiality prediction for the FBA model Yeast8, yielding a Matthew's correlation coefficient (MCC) of 0.48. c. Gene essentiality prediction for the genes expressed in the cEFL model of Yeast8, yielding a MCC of 0.50. d. Gene essentiality prediction for the cETFL model of Yeast8, yielding a MCC of 0.50.	108
3.3	Comparison of experimental data from Van Hoek <i>et al.</i> (135) (crosses) and simulation results (lines) for the Crabtree effect. Absolute exchange rates of glucose (orange), O ₂ (purple), CO ₂ (light blue), and ethanol (dark blue) are shown. a. cEFL model. b. vEFL model. c. cETFL model. d. vETFL model.	110
3.4	Workflow for the integration of enzyme data into the model. The enzyme composition for the complex enzymes was found in YeastCyc (65) and ComplexPortal (150). The function <code>matchKcats.m</code> from GECKO (38) was used to find catalytic rate constants.	118
4.1	Simplified mechanism of action of metformin on a human cell. AMPK: AMP-activated protein kinase K.; PKA: Protein Kinase A; ACCOAC: Acetyl-CoA carboxylase; <i>PCK1</i> : gene of the cytoplasmic phosphoenolpyruvate carboxykinase 1; PYK: Pyruvate kinase	124
4.2	Specific growth rate, flux rates, and enzyme levels in the colon cancer ETFL ME-model, for different metformin concentrations in mM. Fluxes have error bars that represent their variability at fixed growth rate, and the vertical bars represent their rate at the Chebyshev center. a. Specific growth rate of cells as a percentage of the specific growth rate of a cell without metformin treatment (negative control – left bar). The control corresponds to a specific growth rate of 0.034 h ⁻¹ . b. Flux (mmol · g _{DW} ⁻¹ · h ⁻¹) through the NADH:ubiquinone oxidoreductase, Type I NADH dehydrogenase (respiratory complex I). c. Flux (mmol · g _{DW} ⁻¹ · h ⁻¹) through the mitochondrial ATP synthase. d. Concentration of the mitochondrial complex I enzyme, in g/g _{DW}	128
4.3	Reactions of the respiration pathway involving the cofactors NAD/NADH, Q10/Q10H2, FAD/FADH2. The respiratory complex I uses electrons from the cellular NADH pool. The respiratory complex II uses succinate as an electron source. Q10H2 then follows the rest of the respiration pathway through complexes III and IV	129

4.4	Flux rates of reactions involved in the FADH-mediated respiratory process, at the Chebyshev center, for different metformin concentrations in mM. a. Respiratory complex II, transferring electrons from FADH ₂ to Q10. b Succinate dehydrogenase, which transfers electrons to FAD to form FADH ₂	129
4.5	Enzyme levels, and flux rates in the colon cancer ETFL ME-model, for different metformin concentrations in mM. a. AMPK concentration. b. PKA concentration. c. Acetyl-CoA carboxylase flux. d. Concentration of Acetyl-CoA carboxylase, (-) and (+) respectively denote the unphosphorylated and phosphorylated forms.	130
5.1	Variability analysis for reactions whose directions are not constrained by FBA. By subsequently adding thermodynamics constraints and concentration data, all the reaction directionalities are determined.	152
7.1	Different outputs from Skimpy. a. Violin plot of the distribution of the 100 sampled concentrations for the metabolites in the model. b. Euclidean norm of the relative deviation of the 100 different resulting perturbation experiments. Each line represents the norm of the relative change of concentration over time versus the reference concentration. Orange lines converge towards the reference steady state, purple lines towards a new steady state. c. Relative concentration deviation from the reference state, in a specific perturbation experiment, with a highlight on the concentration of glucose 6-phosphate (G6P, light blue), fructose 1,6-biphosphate (FDP, dark blue), ATP (yellow), NADH (orange) and CO ₂ (pink).	166

List of Tables

1.1	Indices, variables, and parameters used in the formulation.	24
1.2	Nomenclature of the models used in the study of <i>E. coli</i> iJO1366. EFL stands for <u>E</u> xpression and <u>F</u> Luxes, ETFL for <u>E</u> xpression, <u>T</u> hermodynamics, and <u>F</u> Luxes, and the v- prefix indicates the inclusion of growth-dependent parameters (see the section Discretization of mRNA and enzyme content)	27
1.3	Properties of the vETFL model generated from iJO1366.	28
1.4	Characteristic completion run times for several types of studies in the vETFL study of iJO1366	39
2.1	Properties of the vETFL model generated from iJO1366.	82
3.1	Properties of the vETFL model generated from Yeast8.	105
3.2	Nomenclature of the models used in the study of the <i>S. cerevisiae</i> model Yeast8. EFL stands for <u>E</u> xpression and <u>F</u> Luxes, ETFL for <u>E</u> xpression, <u>T</u> hermodynamics, and <u>F</u> Luxes. The c- and v- prefixes indicates the inclusion of constant or growth-dependent biomass compositions	106
4.1	Properties of the vEFL model obtained from the colon cancer-specific reduced RECON3D.	126
4.2	Parameter values used in the regulation-enabled ME-model of colon cancer and their sources	141
6.1	Example lumped reactions. 3PG : 3-phospho-D-glycerate; AcCoA : Acetyl-CoA; Asp-L : Aspartate; CoA : Coenzyme-A; CTP : Cytidine triphosphate; DHAP : Dihydroxyacetone phosphate; GLCN : D-gluconate; PE160 : Phosphatidylethanolamine (16:0); PEP : Phosphoenol Pyruvate; Pi : Phosphate; PPi : Pyrophosphate; Q8 : Ubiquinone-8; Q8H₂ : Ubiquinol-8 . . .	159



Introduction

In this first chapter, I introduce the context in which the research I will present is inscribed. I first explain what systems biology is useful for, and the variety of problems that it tackles. Since this thesis is at the boundary between several fields of science, I also introduce important terms and basic language related to the different disciplines connected to my research. Finally, I present the motivations of my research, and the structure of this thesis.

An informal prologue

Systems biology, bioengineering, synthetic biology, computational biology, metabolic engineering: all these terms describe different flavors with which to consider a key scientific problem: understanding and engineering life. Far from Homeric chimeras or Crichton's *Jurassic Park*, understanding and engineering life is a significant challenge of the 21st century. Understanding the human metabolism, and elucidating abnormal processes giving rise to dementia or cancer for instance, are relying heavily on our ability to study these systems with modern tools and methods. Sustainable food production is also an important topic for the science of today and tomorrow, and engineering crops for better resistance to the changing climate will be necessary to alleviate climate-induced economic stress of populations. Finally, the transition from petrochemical to biochemical sourcing of both commodity and specialty chemicals is of paramount importance to minimize the anthropogenic impact on the planet and mitigate the influence of oil and gas geopolitics.

The common anchor between all these topics is their relation to living matter, and cells in particular. Human bodies contain tens of trillions of specialized cells, plants are also multicellular organisms, and microorganisms have long been used for their biochemical capacities, for instance in the fermentation of wine or cheese. A discipline to understand and manipulate cellular behavior is metabolic engineering, a fairly recent field whose inception is conventionally attributed to Prof. James E Bailey in 1991 (1) ². Metabolic engineering consists in altering the functions of a cell to fit a desired objective.

²Which technically makes it a millennial science.

This objective can be the death of metastatic cells, the resistance to a plant virus, or the production of a biochemical of interest. Metabolic engineering uses mathematical models to understand and predict the response of cells to different perturbations. These mathematical models benefit from computational formulations, that help represent *in silico* what happens *in vivo*. The variety of tools and methods developed in that regard span the field of computational biology. Computational biology helps produce engineering designs that can be implemented in living systems through synthetic biology. Synthetic biology also includes the development of tools to engineer these living systems. Because of the intricate nature of biology, successful engineering requires an inclusive view of interactions and mechanisms between the cellular elements and the cellular environment. This holistic view is a defining characteristic of systems biology.

In this thesis, I develop some new, and some improved computational tools for metabolic engineering. In the majority of this work, I use methods from mathematical optimization to model cells at the biochemical level. I model physical, chemical, and biological phenomena with equations and inequalities, and show the resulting models allow to (i) integrate experimental data to characterize observed cellular physiologies; (ii) elucidate non-measurable cellular states underlying observed cellular physiologies; (iii) predict, whenever experimental data are not available, cellular responses that are empirically validated.

★
★ ★

A side note: So are we talking about GMOs? Yes.

When talking to distant relatives, gathering with friends of friends, or meeting new people, it is common to be asked the question “So what is your PhD about?”. An amusing social experiment I have performed was to either answer something along the lines of “I build models to help make biofuels and better understand cancer”, or “I build models to improve GMOs”. As you can imagine, one perspective tended to be judged better than the other, and yet both are technically correct. I think this reveals the important lack of context given to non-specialists with respect to the subject of genetic modification.

High Andean cultures (Aymaras and Quechuas) already had bioengineers 5000 years ago; the selective breeding performed by their farmers allowed them to acclimate crops and livestock for a wide diversity of altitudes, up to 4500m (2). In the 19th century, Gregor Mendel’s records on heredity in pea plants established the first foundations of modern genetics and traits selection (3). The manipulation of the genome of our surrounding living entities has been a part of the human culture since long ago. What changes here is, rather than obtaining the desired changes through a reproduction/selection scheme, modern genome editing techniques allow to directly perform *in vitro* the desired genome

modification.

A full review of genome editing technologies and their impact is outside the scope of this dissertation, and good reviews are available elsewhere (4, 5). However, I believe it is our duty as scientists/engineers to allow the non-specialized population to access factual information about this important controversial technology.

Industrial biotechnology is probably the field where the use of GMOs is the most common. Examples include the industrial bio-production of artemisinic acid, an anti-malarial drug precursor (6), and the production of isoprenoids – chemical compounds of interest for biofuels, bioplastics, and material science (7, 8, 9, 10, 11). Other fields, such as health and agriculture, present a context in which the use of genome editing can be problematic. Genome editing of human cells is one of the most promising technology to treat hereditary diseases, or even HIV (12). Recently, a gene therapy for age-related macular degeneration (13, 14) has been approved for clinical trials. However, gene editing in humans also poses the problem of artificial trait selection (5), and unintended side effects, including the induction of cancer (15). GMO pest resistant crops reduce the use of pesticides and show increased productivity (4, 16). They also out-compete non-GMO crops and reduce the income of potentially smaller, independent producers.

This is but an extremely limited list of the potential applications and drawbacks of the gene editing technologies, which might help in mundane discussions to go beyond bipartisanship and arguments from authority.

★
★ ★

Field-specific context and definitions

The quantitative description of biological systems requires mathematical formulations, and their complexity calls for the use of computational methods (17). Metabolic engineering, computational biology, and systems biology in general, find themselves at the intersection of numerous scientific fields. In this section, I will provide a short digest of several concepts from different disciplines that play a role in this dissertation.

The central dogma of molecular biology

The central dogma of molecular biology is a way to conceptualize the flow of biological information.

Living matter stores information in biopolymers, long molecules made of several of different monomers. All modern living organisms store their genetic information, or

Introduction

genotype, in deoxyribonucleic acid (DNA), a biopolymer composed of nucleotides. There are four types of nucleotides (A,T,C, and G), and their sequence provides a quaternary encoding of the genetic information.

DNA is read (transcribed) by an enzyme (RNA polymerase) into ribonucleic acid (RNA) strands. RNA is also a biopolymer of nucleotides (A, U, C and G), and carries the information to synthesize proteins.

Ribosomes read (translate) triplets of nucleotides into a sequence of amino acids, called a polypeptide. Most organisms use a basis of 21 amino acids³. Because of the physico-chemistry of amino acids, the polypeptides fold and combine each other into proteins. Proteins can be composed of one or several polypeptides (complexes), and the proper folding of the protein will determine its function. The protein profile of a cell is called its proteome, and can be described in terms of concentrations of proteins.

Proteins are critical components to the functioning of cell, and their folding and spatial configuration directly impacts cell physiology. For instance, changing the spatial configuration of proteins is a key regulation mechanism in cells, used to activate or deactivate certain parts of the metabolism. This dependency between shape and activity is also used in anti-cancer treatments and disease resistance mechanisms in a wide range of living organisms. On the other hand, misfolded proteins are involved in the onset Alzheimer's disease, and prion diseases such as Creutzfeld-Jakob disease. Proteins can have a catalytic role, acting as facilitators of biochemical reactions, but also a structural role (for example in sickle cell anemia or bacterial biofilms), or an information-transmission role (signaling cascades, histocompatibility system for recognizing foreign elements in the body). The proteins that catalyze reactions are called enzymes, and control the biotransformations inside the cells, or their metabolism. The metabolism of a cell is responsible for how it grows and evolves in its environment.

The chain of information from the DNA nucleotides to the synthesis of proteins constitutes the gene expression mechanisms of the cell.

Finally, the observable traits of the cell, understood as a product of the information flow from the DNA to the enzymes controlling metabolism, is called the phenotype of the cell.

The processes of DNA, RNA, and protein synthesis can also be altered by small molecules or even RNA and proteins themselves. Proteins can also be altered after they are synthesized, by post-translational modifications (PTMs). These include the binding of small molecules on specific sites that will change the shape and function of the protein.

All these mechanisms in the cell require energy to be performed. An important energy carrier is adenosine triphosphate (ATP), which is made of a nucleoside A (seen in DNA and RNA) and three phosphate groups. By shedding phosphate groups, ATP can communicate

³Some organisms expand it to 23.

energy to reactions, to increase the thermodynamic drive of biotransformations. In the living systems we study here, ATP is generated through the metabolism of carbohydrates (sugars). By oxidizing (burning) carbon compounds, the cell creates a gradient of potential, which energizes the synthesis of ATP. This is called the electron transport chain, and constitutes the source of energy for the living systems we study. Organisms that use molecular oxygen (O_2) to oxidize their carbon sources are called aerobic organisms, and use the respiration pathway converting O_2 into carbon dioxide (CO_2). Some organisms do not need oxygen. For example, *S. cerevisiae* can oxidize its carbon source using other means, at the cost of a reduced efficiency, producing ethanol and CO_2 as a byproduct. Some organisms can also use minerals, and even uranium, to generate energy (18, 19).

Even this summarized view gives a glimpse of how complex, multilevel interactions are important when considering biological problems, and why systems biology came to be.

Constrained optimization

Optimization is a discipline focused on using mathematical tools to find out how a process can be improved to the limit, as well as the study of these limits. Optimization relies heavily on mathematical formulations of the studied process, and in particular on the design of an objective function which translates the desired traits into mathematical properties. Constrained optimization, in particular, defines a problem using variables, or the states of the system, and constraints, or the relationships between variables in the system.

An important fraction of the methods I developed in this work are focused on a special type of constrained optimization: mixed-integer linear programming. Linear programming is a subtype of constrained optimization, where the constraints and objective function can be described as linear (but for real affine) functions of the variables. A typical linear programming problem can be cast in the following manner:

$$\begin{aligned} & \underset{x}{\text{minimize}} && f^\top x \\ & \text{subject to} && a_i^\top x + b_i = 0, && i \in \llbracket 1..N \rrbracket, \\ & && c_j^\top x + d_j \leq 0, && j \in \llbracket 1..M \rrbracket. \end{aligned} \tag{1}$$

Here, a_i, b_i, c_j, d_j, f are real vectors of the dimension of x , our variable. f carries the weights that define preferred components of x for the objective function. For instance, this optimization problem can be used to describe the cost of production of finished goods, where x defines the input and outputs materials, the equalities the non-accumulation of materials, and the inequalities a cap on the material supply. In this context f defines the costs associated to each material.

Mixed integer linear programs are simply linear programs with variables taking integer values, instead of real values. For example, baking a chocolate fondant requires eggs, but one cannot consider putting $5/3^{\text{rd}}$ of an egg in a dough. Integer variables provide a solution subject to the additional constraint that these variables must be integer. Of course, it might mean that the new solution is suboptimal compared to its continuous counterpart (the cake will be costlier, or less good depending on the objective function). Integer variables are also useful to model binary conditions: in our culinary example, if guest allergens are taken as additional variables, we can model the use of only gluten-free flour if one or more guests are gluten-intolerant.

Solving these problems requires a solver, a piece of specialized software that implements specially-designed solving methods. The use of integer variables requires specialized solvers, and comes at a cost in solving speed.

There also exist non-linear solvers, which accept diverse forms of problems that do not present a linear form. Unfortunately, the size of the biological systems we consider in this research is a limiting factor in the use of non-linear solvers, as their complexity may result in impractical long solving times. A part of this thesis discusses the (piecewise-)linearization of a non-linear problem, and how it improves solving speed.

Modeling the cell at the level of metabolism and expression

Genome-scale models

The end of the 20th century saw the advent of whole-genome sequencing. One of the main consequences of the new genomic era was the sudden availability of a wealth of genetic information for different organisms. The genome of the ubiquitous bacterium *E. coli* was sequenced in 1997 (20), that of *S. cerevisiae* (baker's yeast) in 1996 (21), and the Human Genome Project sequenced the first human genome in 2001 (22). The large amount of data available from sequencing promoted the development of new methods to model cells at the genome scale.

In particular, genome-scale models of metabolism (GEMs) that capture different levels of the biology within the cell were developed from the sequencing of organisms. The genetic sequence informs which enzymes, and therefore which reactions, are occurring in the cell. From this information is reconstructed a metabolic network, which describes the biotransformations happening in the cell, in terms of metabolites and reactions. Additionally, the genetic information links most of the reactions to their related genes using gene-protein-reaction association rules (GPRs). GPRs summarize the essentiality of genes in the occurrence of a reaction. A gene may be essential for a reaction to happen, if they code for a peptide that is a part of the only enzyme catalyzing the reaction. A gene may also not be essential to the reaction, if several enzymes (isoenzymes) can catalyze

the reaction. GEMs are also able to describe the growth and energy requirements of the cell, as well as the localization of metabolites and reactions within the different biological compartments (23), such as cytosol or mitochondria.

Genome-scale models have been reconstructed for more than 6000 organisms (24, 25, 26), and are the main component to a wealth of techniques to elucidate metabolism in the cell.

Constraint-based models of metabolism

GEMs hold a wealth of information that can be used to model the metabolism of cells. In particular, flux balance analysis (FBA) is a constraint-based method programming that aims at modeling the fluxes of the biochemical reactions in a system at steady state (27, 28). FBA uses the information from GEMs on the reactions known to happen in a cell, and which metabolite they consume or produce, and formulates a linear programming problem. In this formalism, the variables of the FBA problem will be the net fluxes through each reaction, and the constraints will be that the algebraic sum of the producing and consuming reactions of the internal metabolites is equal to zero. These constraints originate from the steady state assumption, and capture the non-accumulation of metabolic intermediates in the cell. Additional constraints can be set for reactions whose flux is known, either through monitoring of the growth medium, or by advanced techniques such as ^{13}C metabolic flux analysis (^{13}C MFA). Being a linear programming formulation, FBA also assumes an objective function governs the cell. A typical FBA problem has the following form:

$$\begin{aligned} & \underset{v}{\text{maximize}} && f^\top v \\ & \text{subject to} && S^\top v = 0, \\ & && \underline{v} \leq v \leq \bar{v}. \end{aligned} \tag{2}$$

S is the stoichiometric matrix of the model, with each rows representing metabolites, columns reactions, and each element the stoichiometric coefficient of a metabolite in a reaction. v is the variable for which we are solving, and represents the reaction fluxes. \underline{v} and \bar{v} represent boundaries on the reactions rates. Finally, f defines the objective function. For instance, it has been shown the growth rates values predicted by FBA in an *E. coli* model that adopts growth maximization as its objective match experimentally measured growth rates (27, 29). Other objective functions that have been considered include the minimization of ATP usage, or the minimization of the sum of fluxes (30), as these objectives are associated to a reduced energy consumption by the cell.

It is important to note that the assumption that the cell physiology follows an objective is reliant on the idea that the cell adopted this specific objective under selective pressure (31,

32). Increased growth rate and reduced energy consumption, are competitive advantages that favor the reproduction of single-celled organisms. However, these assumptions do not always translate easily to more complex systems, such as multicellular organisms, host-pathogen interaction systems, engineered microorganisms, or tumor cells; alternative objective functions should be considered depending on the problem studied, as detailed by Zomorodi *et al.* (30).

Towards more biochemistry in the constraints

Due to its simplicity of execution, FBA has been widely adopted in metabolic engineering, synthetic biology, and systems biology. Multiple methods have also been developed to supplement it with more constraints, capturing physical, chemical, and biological dependencies in the modeled organisms (33, 34). These additional constraints can both improve the models ability to describe cell physiology, and allow to integrate more data in models.

One noteworthy example is thermodynamics-based flux analysis (TFA) (35, 36), which aims at adding thermodynamics constraints to enforce thermodynamically consistent reaction directionalities. The method uses a mixed-integer formulation to enforce the coupling between Gibbs free energy of a reaction (its thermodynamic drive), and the sign of its net flux, or directionality. A reaction with a negative Gibbs free energy will have a positive net flux, and vice versa. Because the TFA formulation introduces metabolic concentrations as variables in the model, it also allows to integrate measurements of the concentration of metabolites (metabolomic data), if it is available. TFA is explained briefly in the methods of Chapter 1, and Chapter 5.

Another direction in which constraint-based models of metabolism have evolved is the modeling of the biological layers underlying metabolism. In particular, resource balance models add a total protein capacity constraint, as formulated in Beg *et al.* (37), and were adapted to integrate protein concentration data (proteomics) by Sanchez *et al.* (38). This addition allows to model the competition for a new resource: since a cell has a limited size, it can only contain a limited amount of proteins – the proteome space is limited. Modeling proteome limitation allows to model new physiologies, such as overflow metabolism or diauxie. Overflow metabolism in *S. cerevisiae* is discussed extensively in Chapter 3, and diauxie in *E. coli* in Chapter 2.

Even below the proteome layer, modeling the whole expression mechanisms of the cell has garnered a lot of attention, giving rise to models of metabolism and expression (ME-models) (39, 40). ME-Models, and their more recent refinement by Lloyd *et al.* (41) were the first effort to integrate from the bottom-up the totality of the expression mechanisms in the cell, from DNA replication to mRNA and protein synthesis, at the genome scale. Since these models introduce variables to represent protein and RNA

abundances, they also allow direct integration of proteomic and transcriptomic data. Transcriptomic data are usually experimentally easier to obtain than proteomics, and ME-models are a valuable tool in this respect, since their integrated formulation helps elucidating genotype-phenotype relationships. However, the macromolecules (DNA, RNA, proteins) introduce a non-linear term in the constraints of the optimization problem, making it more challenging to solve. This topic is discussed in detail in Chapter 1.

Formulating together TFA and ME-models approaches holds the promise of a strong, more holistic framework for multi-omics integration, as well as a robust method to explore genotype-phenotype relationships. However, the combination of mixed-integer problems, the non-linearity of ME-models, and the sheer size of the obtained models, make the joint formulation a challenging endeavor. This will be the main topic of this dissertation.

Motivations

The fast rate of emergence of metabolic engineering applications, from health and food to biochemicals, emphasizes the necessity to develop new models and methods to fuel the fast leaps forwards the field must achieve to address the challenges of the 21th century (17, 42).

The first main theme of this thesis is to improve the current state of the art on models of metabolism and expression, and provide new, more efficient methods to understand and engineer cells in various contexts. In particular, providing new methods for data integration is of paramount importance to improve both the interpretation of experimental data and the accuracy of models. These models also pave the way to cell-specific modeling, which is a keystone part of both industrial biotech and personalized medicine, and even constitute a significant step towards whole-cell models, an overarching aim of systems biology.

A second main theme in this work is to use the developed method to reproduce non-trivial phenotypes, in agreement with experimental data, and with a minimal set of assumptions. Moreover, understanding the emergence of genetic control on the cell metabolism as a product of the evolutionary pressure on a system will be key to deciphering complex cellular behavior, and ultimately engineering them. Specific examples include the interpretation of overflow metabolism as an optimal behavior under resource constraints, or the emergence of robust cellular control strategies to ensure growth optimality.

Both these themes further demonstrate the utility and predictive power of models. Many variables of the model are not measurable in real systems or in real time, and the model can fill the lack of data and provide insights on how these hidden variables behave. The study of the feasible inner states of the system will also help implement successful engineering strategies, either to improve the productivity of a biochemical production

process, or the design of drug strategies against diseases.

Structure of the thesis

This thesis comprises two parts and seven chapters.

The first part focuses on the development of a new ME-model formulation that includes thermodynamics, and its applications in several organisms. New constraint-based methods are also detailed to simulate complex phenotypes.

Chapter 1 details a new MILP-based formulation of ME-models, called ETFL. The formulation is then applied to an *E. coli* model, and used to simulate proteome-limited growth. Several types of constraints are introduced, both in the general formulation and for more specific cases.

Chapter 2 presents a dynamic formulation of cellular growth in a batch reactor, using ETFL. The formulation is able to test and confirm our hypothesis that the phenomenon of diauxie is explained at the proteome level, and account for different cell fates depending on initial conditions. The formulation also introduces the use of Chebyshev centering for finding a robust representative of the ME-model solution space.

Chapter 3 shows the appearance of overflow metabolism and ethanol production in aerobic cultures of *S. cerevisiae*, or Crabtree effect, is a consequence of optimal proteome-limited growth. For this purpose, ETFL was used to produce the first ME-models of yeast, and in general, of a eukaryotic organism.

Chapter 4 demonstrates the further uses of the ETFL framework to study cell-drug interactions and signaling cascades. In particular, a ME-model generated from a tissue-specific reduced model of a human colon cancer cell line is developed, and the impact of the antidiabetic drug metformin on tumor growth is reproduced. We also show how such a framework can be used to infer the emergence of resistance mechanisms to the drug, and highlight two different growth-limiting actions of metformin that occur at different doses.

The second part of the thesis consists on short chapters presenting software that I developed and contributed to. All this software is openly accessible on online repositories.

Chapter 5 describes an implementation of TFA in Python that I used in all the studies described in this work.

Chapter 6 describes an algorithm to systematically reduce genome scale models around subsystems of interest, based on pyTFA.

Chapter 7 describes a software package to generate and analyse symbolic kinetic models.

These models can be automatically imported from pyTFA, and the package automatically generates the expressions to perform, for instance, metabolic control analysis, full time integration, or sensitivity analysis.

Finally, I conclude this dissertation with summarizing remarks and future perspectives of this work.

Integrating expression mechanisms in genome-scale models

I

We mathematicians are all a bit crazy.
— Lev Landau


La science est surtout
une prise de conscience de plus en plus complète
de ce qui peut et doit être découvert.
— Boris Vian

A major part of the research I present in this thesis is related to better modeling cell systems, using optimization. ME-models have been out for some time, and I even had the chance to work for some time prior to my PhD with Joshua Lerman, who with Edward O’Brien authored the first papers on ME-models. Yet, the adoption of ME-models is not as widespread as that of genome-scale models of metabolism. Three factors explain the limited adoption of ME-models as compared to metabolic models: (i) the overall increased complexity of the models; (ii) their increased solving times; and (iii) the lack of integration to mainstream pipelines. Item (i) is quite difficult to tackle, since in the end there are only so many ways to describe the gene expression machinery. (ii) is also a problem that I could not solve alone, as developing my own solver was well beyond the scope of my thesis. (iii), however, was an issue within my technical reach.

Hence, to two initial motivations for which I developed this new ME-model formulation were to make ME-models more accessible, and integrate them with other strong frameworks, in particular thermodynamics-based flux analysis, a trademark technique from my

laboratory. Thus ETFL was born. By developing ETFL, I managed to deconvolute the complicated formulation of models of gene expression and metabolism into smaller, simpler elements, with a direct connection to the biochemistry of the cell. Additionally, the transformation of the formulation we propose, using scaling methods and mixed-integer programming, results in an improvement in the time needed to solve ME-models. To some extent, ETFL managed to tackle the three points mentioned above, which I believe will help make ME-models more accessible and transparent.

I applied ETFL to *E. coli*, *S. cerevisiae*, and even to a context-specific human model. Along the way, I also adapted the formulation to model more than simply gene expression and metabolism, devising a dynamic extension of ETFL and formulating methods to model cell signaling interactions. Because of this, I like to think of ETFL as a new and improved version of ME-models.



1. The ETFL formulation allows
multi-omics integration in
thermodynamics-compliant
metabolism and expression models

Chapter 1. The ETFL formulation allows multi-omics integration in thermodynamics-compliant metabolism and expression models

Pierre Salvy¹, Vassily Hatzimanikatis^{1,*},

¹ Laboratory of Computational Systems Biotechnology, École Polytechnique Fédérale de Lausanne (EPFL), Lausanne, Switzerland

* Corresponding author: vassily.hatzimanikatis@epfl.ch

This first chapter details a new way of formulating genome-scale models of metabolism and expression (ME-Models) for organisms. Indeed, efficiently accounting for mRNAs and enzyme expression in genome-scale metabolic models has been challenging. Here, I introduce a model formulation that simulates thermodynamic-compliant fluxes and enzyme and mRNA concentration levels, that is more efficient than the previous state of the art, and requires less specialized solving methods.

The chapter is adapted from P. Salvy and V. Hatzimanikatis, “The etfl formulation allows multi-omics integration in thermodynamics-compliant metabolism and expression models,” *Nature Communications*, vol. 11, no. 1, pp. 1–17, 2020. Vassily Hatzimanikatis and I worked on the formulation, problem scaling, and designed the studies to perform. I wrote all the code to implement the formulation, and the scripts to perform the studies. I curated the data to make the models, made the figures, and set up the online code repository. The latter includes a documentation I wrote, a continuous integration system to verify the code portability, and several tutorials to reproduce the results.

All the code and documentation is available under the APACHE 2 license at:

<https://github.com/EPFL-LCSB/etfl>

<https://gitlab.com/EPFL-LCSB/etfl>

The content of this chapter is partially reproduced from the original article, with the authorization from the publisher following the guidelines at <https://www.nature.com/nature-research/reprints-and-permissions/permissions-requests> as found on June 8th, 2020. In particular, it is specified: “*Authors have the right to reuse their article’s Version of Record, in whole or in part, in their own thesis. Additionally, they may reproduce and make available their thesis, including Springer Nature content, as required by their awarding academic institution.*”

1.1 Introduction

Metabolic modeling, which helps making sense of the metabolism in a biological network, is an important tool for engineering biocatalysts, with applications in biofuels, drug design, microbial community analysis, and personalized medicine. Model accuracy is instrumental to the success of these applications through an efficient engineering of the host organisms. However, incorporating expression information into metabolic networks poses a significant challenge, and most current models do not even attempt it—effectively excluding an important network in biological systems that can drastically affect results. In metabolic engineering, strains are modified and controlled at the genome level through the transcriptome, and the effects are observed at the fluxome level, which accounts for the range of metabolic reactions in an organism. In between these two levels is the proteome that performs the biochemical transformations according to the genetic template, though it is this middle step in the process that cannot yet be robustly and efficiently incorporated into models of metabolic systems. Because of the complex interplay between these different layers of control, understanding expression and incorporating this into future models is key for improving metabolic engineering.

Classically, model-based strain design has relied on tools that use the DNA sequence of an organism and homology with well-studied organisms to infer a network of metabolic reactions that happen inside a cell of that organism, which is called a genome-scale model (GEM). With current technologies and tools like metagenome sequencing (44), it is possible to generate GEMs for hundreds of different species at a time. GEMs are particularly amenable to flux balance analysis (FBA), which models metabolism at the fluxome level using linear optimization techniques. However, plain FBA has been known to predict biochemically unrealistic solutions like free high-flux cycles or thermodynamically infeasible pathways. It also scales growth linearly with carbon uptake, which is not observed at high-uptake fluxes. FBA also fails to capture growth-dependent and protein-level effects, such as enzyme saturation or proteome-related limitations. Hence, several efforts have been made to supplement FBA with additional constraints to improve its predictive power. For example, thermodynamics-based flux analysis (TFA) (35, 36) uses thermodynamic constraints to enforce thermodynamically consistent reaction directionalities and to allow the integration of metabolomics. Resource balance models add a total proteome capacity constraint, as formulated in Beg *et al.* (37), to model the proteome-related limitations of the cell, as enzymes have to compete for the constrained total amount of cellular proteins. Frameworks like GECKO (38) further build on this resource balance idea and include flux constraints based on proteomics, such as $v \leq V_{max} = k_{cat} [E]$ as well as a constraint on the total proteome mass. Finally, metabolomics and expression models (ME-models) (39, 40) were the first to integrate the entirety of the expression mechanisms of the cell from the bottom-up, including mRNA and protein synthesis.

However, simultaneously accounting for all of these constraints is challenging because

Chapter 1. The ETFL formulation allows multi-omics integration in thermodynamics-compliant metabolism and expression models

of the formulation of each method, as TFA models involve integer variables that yield a mixed-integer linear program (MILP), whereas ME-models involve bilinear constraints that require special optimization procedures and a high-precision (quad-precision) solver (41, 45, 46). Mixing these methods would require the inclusion of integers in ME-models, which is not straightforward and would lead to more complex mixed-integer non-linear programs (MINLP) that are computationally intensive to solve. Additionally, the amount of RNA and protein, the RNA and protein expression rates, and their stabilities are all growth dependent (47), and including accurate representations of these variables leads to even more complex, non-linear models. Meanwhile, although resource balance models such as GECKO could theoretically be integrated into TFA or ME-models in the current formulations, to the best of our knowledge, no link with TFA or ME-models has been proposed. Therefore, the metabolic engineering community needs a common formulation for these methodologies to build the most accurate models.

We investigated the development of such a framework and propose herein a unified formulation for Expression and Thermodynamics-enabled FLux models (ETFL) that can account for the above integration issues. To our knowledge, ETFL is the first formulation that can account at the same time for expression, thermodynamics, and growth-dependant variables. It is also the first to do so using common double-precision MILP solvers. In ETFL, we address the compatibility of the formulations by expressing the growth rate variable in bilinear products as a piece-wise constant function. We also address the issue of solver precision by performing a scaling that reduces the range of orders of magnitude of the variables. This reformulation allows us to transform the problem into a MILP, which we can solve efficiently using common open source or commercial solvers. The resulting model is then effectively able to directly integrate thermodynamic constraints as well as expression constraints and growth-dependent parameters. In this model, metabolite, enzyme, and mRNA concentration levels are explicitly defined to enable fast and easy omics integration: metabolites through their log-concentration variables in thermodynamics constraints, and enzymes and mRNA through their total concentration variables in the expression constraint. Finally, we show an application of this framework to a well characterized *E. coli* model, iJO1366 (48).

Important assumptions are made to derive this formulation. The two most notables ones are (i) we can neglect the dilution rate of metabolites, and (ii) the steady state approximation holds. While these assumptions are commonly made in FBA, we discuss them in details in the Supplementary Note 3, where we also assess their validity in a context where macromolecules are taken into account. Briefly, these assumptions hold because (i) the dilution rate of the metabolites is negligible in front of their synthesis and consumption rates, and (ii) the dynamics of metabolism (including expression) are faster than that of the environment of the cell.

1.2 Results and Discussion

1.2.1 Formulation of the expression problem

ETFL is an ME-model implementation because it proposes a formulation that both accounts for metabolism and expression constraints. ME-models do not aim to replace kinetic models, but to account for the expression cost of making the enzymes that are necessary to carry a biochemical flux. In ETFL, this includes the cost of peptide and mRNA synthesis, as well as the competition for ribosomes and RNA polymerase in a limited proteome.

To transparently account for expression mechanisms and increase the predictive power of our models, we needed to derive the equations that could bridge the biochemistry with the optimization problem that is ETFL. Here we present a summary of these equations, and detail their derivation in the section Materials and Methods. We derived these equations using assumptions similar to those used in the formulation of the GECKO (38) and ME-model (39, 40, 41).

This formulation relies on derivations rooted in the biological mechanism of expression and depends on a number of biochemical parameters related to the cell. In particular, the mass balances of the macromolecules are expressed using concentration variables. Each mass balance will yield an equation where the concentrations of the macromolecules will be variables, thus effectively formulating a new constraint of the model and allowing us to calculate concentration values by solving the model.

We can write the quasi-steady state mass balance for macromolecules as follows:

$$v_{\star}^{\text{syn}} - v_{\star}^{\text{deg}} - \mu * G_{\star} = 0, \quad (1.1)$$

where \star represents the indexing of the macromolecule, v_{\star}^{syn} is the synthesis term, v_{\star}^{deg} is the degradation term, and $\mu * G_{\star}$ is the dilution term. The asterisk “ $*$ ” signifies the product of two variables. The detail of the derivation is available in the Materials and Methods.

Using this formalism, for each macromolecule we can define and link together a synthesis flux, a degradation flux, and the macromolecule’s concentration. Knowing enzyme concentrations allows us to bound the variables representing metabolic reaction fluxes with their maximum catalytic rate according to the classical equation:

$$v \leq k_{\text{cat}} \cdot E, \quad (1.2)$$

where k_{cat} is the catalytic rate constant of the enzyme E with respect to flux v . The dot product “ \cdot ” signifies here a product between a parameter value and a variable. In this same fashion, we can also constrain the synthesis flux for the peptides, which are then assembled into enzymes. Peptide synthesis is simply a metabolic reaction that consumes energy (under the form of GTP) and charged tRNAs and produces a peptide and uncharged tRNAs. The catalytic rate of the reaction is proportional to the maximum ribosomal catalytic rate divided by the length of the peptide to be synthesized. The same can be said about mRNA synthesis, which uses nucleoside triphosphates and is catalyzed by the RNA polymerase. The constraints are explained in the Materials and Methods, in which we detail a *de novo* derivation of the constraint set that describe the expression problem.

The part of the matrix that has been added to the FBA problem to account for expression has been termed the expression problem (EP). Although this initial formulation is bilinear, we detail in the Materials and Methods section how we cast it to a MILP.

Biomass reaction synthesis and mass balance In FBA, the biomass reaction is an artificial, lumped reaction that represents the consumption of metabolites in proportion to the cell growth rate. This consumption reflects nucleoside triphosphate (NTP) requirements for mRNAs, amino acid requirements for proteins, lipid requirements for the cell wall, or metal ion needs. Biomass reaction inclusiveness depends on the modeling assumptions made during the model curation process and can vary significantly among models of the same species. The consumed amount of each metabolite is usually estimated experimentally by measuring the amounts of these metabolites in dried cell mass. Because the stoichiometric ratios of metabolites in the biomass reaction are fixed, the abundance of metabolites is the same for all growth rates. This simplifying assumption, necessary in FBA, goes against experimental evidence. Neidhardt and Curtis (47) report for instance that mRNA and protein mass ratios in the cell change with growth rate.

Because ETFL has explicit expression requirements through transcription, translation, and tRNA-charging reactions, it is possible to account for varying ratios of NTPs and amino acids as the growth rate changes, an effect that is captured in experiments (47). In this context, the approximation made in FBA can be written using ETFL terms:

$$\forall aa_i, \quad \eta_{aa_i}^{v_{\text{biomass}}} \cdot \mu \approx v_{aa_i}^{\text{charging}}, \quad (1.3)$$

$$\forall \text{NTP}, N \in \{T; C; G\}, \quad \eta_{\text{NTP}_i}^{v_{\text{biomass}}} \cdot \mu \approx \sum_{j \in \mathcal{J}} v_{\text{NTP}_i}^{\text{tcr}_j}, \quad (1.4)$$

$$\eta_{\text{ATP}}^{v_{\text{biomass}}} \cdot \mu \approx \sum_{j \in \mathcal{J}} v_{\text{ATP}_i}^{\text{tcr}_j} + v_{\text{ATP}}^{\text{GAM}'}, \quad (1.5)$$

where v^{biomass} represents the biomass equation, and $\eta_{m_i}^{v^{\text{biomass}}}$ is the participation of metabolite m_i in the biomass reaction. $v_{\text{ATP}}^{\text{GAM}'}$ is the growth-associated ATP maintenance that is not linked to expression mechanisms. This includes, for example, phosphorylation requirements (if not modeled in ETFL), polysaccharide synthesis, cofactor regeneration in the biomass reaction, unmodeled organelle functions, and other ATP-hydrolyzing events in the cell⁴.

For each metabolite participating in the biomass reaction, its associated expression is obtained by equating the mass balance constraints in ETFL and in FBA. Hence, to avoid accounting for the expression requirements twice (once through the biomass equation, once through the EP), it is necessary to remove the participation of these metabolites linked to expression from the biomass reaction.

Summary of the formulation Here we show the formulation of the constraints of ETFL. For clarity, we use different indexing sets, each referring to a specific object in the model. The definition of these, as well as that of the variables and the parameters, are detailed in Table 1.1. The formulation of the following equations and an explanation of the specific cases for RNA polymerase and ribosomes is discussed in details in the section Materials and Methods.

⁴This paragraph and Eq. 1.5 have been added to clarify the original text from the published article, which did not explicitly mention non-expression related growth-associated ATP

Chapter 1. The ETFL formulation allows multi-omics integration in thermodynamics-compliant metabolism and expression models

Table 1.1. Indices, variables, and parameters used in the formulation.

Index letter	type	Refers to	Set or unit
i	index	Metabolite	\mathcal{I}
aa_i	index	Amino acid	\mathcal{A}
j	index	Reaction/Flux/Enzyme	\mathcal{J}
l	index	Gene/Peptide/mRNA	\mathcal{L}
s	index	Binary coefficient for growth discretization	$\mathcal{S} = \{0..\lceil \log_2 N \rceil\}$
u	index	Binary coefficient for interpolation discretization	$\mathcal{U} = \{0..N\}$
μ	variable	Growth rate	h^{-1}
v_j^\pm	variable	j^{th} net positive/negative biochemical flux	$\text{mmol.gDW}^{-1}.\text{h}^{-1}$
E_j	variable	Concentration of the j^{th} enzyme	mmol.gDW^{-1}
F_l	variable	Concentration of the l^{th} mRNA	mmol.gDW^{-1}
P_l	variable	Concentration of the RNA polymerase assigned to the l^{th} mRNA	mmol.gDW^{-1}
R_l	variable	Concentration of the ribosome assigned to the l^{th} peptide	mmol.gDW^{-1}
$T_{aa_i}^u$	variable	Concentration of the i^{th} uncharged tRNA	mmol.gDW^{-1}
$T_{aa_i}^c$	variable	Concentration of the i^{th} charged tRNA	mmol.gDW^{-1}
v_l^{tsl}	variable	Translation rate of the l^{th} gene	$\text{mmol.gDW}^{-1}.\text{h}^{-1}$
v_l^{tcr}	variable	Transcription rate of the l^{th} gene	$\text{mmol.gDW}^{-1}.\text{h}^{-1}$
v_j^{asm}	variable	Assembly rate of the j^{th} enzyme	$\text{mmol.gDW}^{-1}.\text{h}^{-1}$
v_j^{deg}	variable	Degradation rate of the j^{th} enzyme	$\text{mmol.gDW}^{-1}.\text{h}^{-1}$
v_l^{deg}	variable	Degradation rate of the l^{th} mRNA	$\text{mmol.gDW}^{-1}.\text{h}^{-1}$
$v_{aa_i}^{\text{charging}}$	variable	Charging rate of the i^{th} tRNA	$\text{mmol.gDW}^{-1}.\text{h}^{-1}$
$k_{\text{cat}}^{j,\pm}$	parameter	Forward/backward catalytic rate constant of the j^{th} net biochemical flux	h^{-1}
k_{deg}^j	parameter	Degradation rate constant of the j^{th} enzyme	h^{-1}
k_{deg}^l	parameter	Degradation rate constant of the l^{th} mRNA	h^{-1}
η_l^j	parameter	Stoichiometry of the l^{th} peptide in the j^{th} enzyme	$[\emptyset]$
$\eta_{aa_i}^l$	parameter	Stoichiometry of the i^{th} amino acid in the l^{th} peptide	$[\emptyset]$
L_l^{aa}	parameter	Length in amino acids (aa) of the l^{th} peptide	aa
L_l^{nt}	parameter	Length in nucleotides (nt) of the l^{th} mRNA	b
$L_{\text{rib}}^{\text{nt}}$	parameter	Ribosome footprint size on mRNA, in nucleotides	b
ρ	parameter	Ribosome occupancy	$[\emptyset]$
π	parameter	RNA polymerase occupancy	$[\emptyset]$

Metabolite mass balance

$$S \cdot v = 0 \quad (\text{FBA})$$

Catalytic constraints

$$v_j^+ - k_{\text{cat}}^{j,+} E_j \leq 0 \quad (\text{FC}_j)$$

$$v_j^- - k_{\text{cat}}^{j,-} E_j \leq 0 \quad (\text{BC}_j)$$

Expression mass balance

$$v_l^{\text{tsl}} - \sum_{j \in \mathcal{J}} \eta_l^j \cdot v_j^{\text{asm}} = 0 \quad (\text{PB}_l)$$

$$v_{\text{rRNA}_l}^{\text{tr}} - v_{\text{rib}}^{\text{asm}} = 0 \quad (\text{RB}_{\text{rRNA}_l})$$

$$v_j^{\text{asm}} - v_j^{\text{deg}} - \mu * E_j = 0 \quad (\text{EB}_j)$$

$$v_l^{\text{tr}} - v_l^{\text{deg}} - \mu * F_l = 0 \quad (\text{MB}_l)$$

$$-v_{\text{aa}_i}^{\text{charging}} + \sum_{l \in \mathcal{L}} \eta_{\text{aa}_i}^l \cdot v_l^{\text{tsl}} - \mu * T_{\text{aa}_i}^u = 0 \quad (\text{TB}_{\text{aa}_i}^u)$$

$$v_{\text{aa}_i}^{\text{charging}} - \sum_{l \in \mathcal{L}} \eta_{\text{aa}_i}^l \cdot v_l^{\text{tsl}} - \mu * T_{\text{aa}_i}^c = 0 \quad (\text{TB}_{\text{aa}_i}^c)$$

$$v_{\text{DNA}}^{\text{synthesis}} - v_{\text{DNA}}^{\text{deg}} - \mu * \text{DNA} = 0 \quad (\text{DB}_{\text{DNA}})$$

Degradation fluxes

$$v_j^{\text{deg}} - k_{\text{deg}}^j \cdot E_j = 0 \quad (\text{ED}_j)$$

$$v_l^{\text{deg}} - k_{\text{deg}}^l \cdot F_l = 0 \quad (\text{MD}_l)$$

Expression constraints

$$v_l^{\text{tr}} - \frac{k_{\text{cat}}^{\text{RNAP}}}{L_l^{\text{nt}}} P_l \leq 0 \quad (\text{TR1}_l)$$

$$v_l^{\text{tsl}} - \frac{k_{\text{cat}}^{\text{rib}}}{L_l^{\text{aa}}} R_l \leq 0 \quad (\text{TR2}_l)$$

$$R_l - \frac{L_l^{\text{nt}}}{L_{\text{rib}}^{\text{nt}}} F_l \leq 0 \quad (\text{EX}_l)$$

$$P_l - \frac{L_l^{\text{nt}}}{L_{\text{RNAP}}^{\text{nt}}} \cdot n_l \cdot \text{DNA} \leq 0. \quad (\text{CN}_l)$$

Chapter 1. The ETFL formulation allows multi-omics integration in thermodynamics-compliant metabolism and expression models

Total capacity

$$\sum_{l \in \mathcal{L}} R_l + R_F - E_{\text{rib}} = 0 \quad (\text{TC2})$$

$$\sum_{l \in \mathcal{L}} P_l + P_F - E_{\text{RNAP}} = 0 \quad (\text{TC1})$$

$$R_F - (1 - \rho) E_{\text{rib}} = 0 \quad (\text{RR})$$

$$P_F - (1 - \pi) E_{\text{RNAP}} = 0 \quad (\text{PR})$$

Macromolecule fractions

$$\sum_{j \in \mathcal{J}} \text{MW}_j \cdot E_j - \sum_{u \in \mathcal{U}} \lambda_u \cdot P^m(\mu) = 0 \quad (\text{IC}_{\text{Enz}})$$

$$\sum_{l \in \mathcal{L}} \text{MW}_l \cdot F_l - \sum_{u \in \mathcal{U}} \lambda_u \cdot R^m(\mu) = 0 \quad (\text{IC}_{\text{mRNA}})$$

$$\text{MW}_{\text{DNA}} \cdot \text{DNA} - D^m(\mu) = 0 \quad (\text{IC}_{\text{DNA}})$$

DNA Synthesis

$$v_{\text{DNA}}^{\text{synthesis}} - \frac{k_{\text{pol}}^{\text{DNAPol3}}}{L_l^{\text{bp}}} \text{DNAPol3} \leq 0 \quad (\text{DP})$$

Recovering the FBA problem In the ETFL formulation, enzyme synthesis is driven by the coupling between FBA and EP through the catalytic constraints. To carry flux, the cell needs to produce enzymes whose production will also use the metabolic resources of the cell. If allocation constraints are enforced, the amount of protein and mRNA synthesized must meet predefined mass ratios for the problem to be feasible. Hence, the metabolic requirement terms for the expression machinery (amino acids and NTP) have been removed from the biomass reaction and are accounted for in the tRNA charging and transcription reactions. Thus, the FBA solutions can be recovered from the ETFL formulation by the following routine:

- Setting $\forall j, \quad k_{\text{cat}}^{j,\pm} = +\infty,$
- Constraining $\forall \text{aa}_i, \quad v_{\text{aa}_i}^{\text{charging}} = \eta_{\text{aa}_i}^{v_{\text{biomass}}} \cdot \mu,$
- Constraining $\forall \text{NTP}_i, \quad \sum_{l \in \mathcal{L}} v_{\text{NTP}_i}^{\text{tr}_l} = \eta_{\text{NTP}_i}^{v_{\text{biomass}}} \cdot \mu,$
- If applicable, relaxing the allocation constraints,
- If applicable, relaxing the thermodynamic coupling constraints.

Table 1.2. Nomenclature of the models used in the study of *E. coli* iJO1366. EFL stands for Expression and Fluxes, ETFL for Expression, Thermodynamics, and Fluxes, and the v- prefix indicates the inclusion of growth-dependent parameters (see the section Discretization of mRNA and enzyme content)

	growth-independent parameters	growth-dependent parameters
(-) thermodynamics	EFL	vEFL
(+) thermodynamics	ETFL	vETFL

1.2.2 Application: *E. coli* genome-scale model iJO1366

iJO366 (48) is a well-curated and well-studied GEM of *E. coli* that is closely related to the GEM used in developing both ME-models iOL1650-ME (40) and iJL1678b-ME (41). Additionally, this model has been extensively applied in the literature and is aligned with a variety of datasets that can be used for data integration. We wanted to subject the model to classical studies that would highlight the power of ETFL, particularly as pertains to proteome-limited growth, macromolecule concentration variability analysis, and gene knock-out studies. We also wanted to assess the sensitivity of the model with respect to the presence of thermodynamic constraints as well as growth-dependent parameters.

Thus, we first experimented with four different models using ETFL with or without thermodynamic constraints and growth-dependent protein/RNA/DNA allocation following Table 2 as reported by Neidhardt *et al.* (47). The following Table 1.2 details the nomenclature used to refer to these different models. The features of the most constrained model containing both thermodynamic and growth-dependent parameters, vETFL, are detailed in Table 1.3. These four models were optimized for maximal growth at increasing glucose uptake rates to assess their behavior with respect to excess substrate, which will show the non-linearity of the relationship between growth and glucose uptake at high uptake rates. A plateau in the growth rate was expected, which indicates a proteome-limited phenotype that cannot be observed with FBA. We also subsequently subject vETFL to a variability analysis and gene essentiality analysis, which will respectively show us the flexibility of the model and its accuracy in predicting gene knock-out behavior.

Growth rate prediction To study the behavior of the model at different carbon uptake rates, we simulated growth on a minimal medium with only glucose as a carbon source, unlimited oxygen, and some essential inorganic compounds. This would allow us to show that at a higher carbon uptake, the model would predict a limited growth – unlike FBA that would predict an unlimited linear increase.

Figure 1.1⁵ shows the predicted growth rate of the different (v)E(T)FL models described in Table 1.2 with respect to the glucose uptake of the cell. As expected and in contrast to current FBA models, all four models plateau after a certain uptake rate, which indicates

⁵This figure has been updated with the latest models for this thesis. As such, it differs from the figure that was published in the original article.

Table 1.3. Properties of the vETFL model generated from iJO1366.

Growth upper bound $\hat{\mu}$	$3.5h^{-1}$
Number of bins N	128
Resolution $\frac{\hat{\mu}}{N}$	$0.0273h^{-1}$
Number of constraints	42992
Number of variables	33923
Number of species	3367
– Metabolites	1809
– Enzymes	563
– Peptides	475
– mRNA	475
– rRNA	3
– tRNA	21×2
Number of reactions	5157
– Metabolic	1542
– Transport	740
– Exchange flux	324
– Transcription	475
– Translation	475
– Complexation	563
– Degradation	1038
Number of metabolites $\Delta_f G'^o$	1558
Number of reactions $\Delta_r G'^o$	1786
Percent of metabolites $\Delta_f G'^o$	86.1%
Percent of reactions $\Delta_r G'^o$	78.3%

a proteome-limited phenotype due to the limited capacity of the cells to make more enzymes to metabolize the glucose. As discussed for the ME-models (40) and GECKO (38) formulations, within the context of models accounting for protein usage, this is caused by (i) the protein burden necessary to metabolize higher fluxes; (ii) the increased demand in protein synthesis at higher growth rates; and (iii) for the models with allocation constraints, the allowed protein and RNA mass ratio. We can see that models featuring protein, RNA, and DNA allocation constraints (vE[T]FL) consistently predict a lower growth rate than models without allocation constraints. This is expected, as the data we input requires additional proteins and mRNA to account for non-metabolism-related macromolecules. Models featuring thermodynamic constraints ([v]ETFL) also predict a lower growth rate, consistent with the fact that thermodynamic constrain the model to valid solutions whose flux is in the subspace of the FBA feasible space. The most constrained model (vETFL) consequently has the lowest growth rate at any glucose uptake. This is in accordance with published TFA results that eliminated biologically infeasible flux profiles yielding non-realistic higher growth rates (35).

We summarize the constraint matrix of the EP of vETFL in Supplementary Table 1, where each line represents a type of constraint and each column represents a type of variable. The blocks of the matrix that are non-zero are colored, and these blocks directly reflect the involvement of the constrained variables.

Modeling missing enzymes Although we initially focused on including only enzymes for which we had all the necessary information (catalytic rate and peptide constitution), we wanted to assess the robustness of our model when the missing enzymes were modeled as well as check our model’s sensitivity to changes in the catalytic rate constants. Thus, we additionally built three more models, based on vETFL, with the following properties: (i) all the missing enzymes were estimated by averaging the properties of the known enzymes based on the curation for the vETFL iJO1366 (333 amino acids long, average $k_{\text{cat}}^{\pm} = 172 \text{ s}^{-1}$); (ii) all the enzymes (including the missing enzymes) but the ribosome, RNA polymerase, and ATP synthase were assumed to have an average catalytic rate constant $k_{\text{cat}}^{\pm} = 172 \text{ s}^{-1}$; and, for comparison purposes, (iii) all the known enzymes of vETFL except for the ribosome, RNA polymerase, and ATP synthase were assumed to have an average catalytic rate constant of $k_{\text{cat}}^{\pm} = 172 \text{ s}^{-1}$. For clarity, we will refer to these models as (i) the model with estimated enzymes; (ii) the all-average model; and (iii) the partial-average model. The ribosome, RNA polymerase, and ATP synthase were not modified, as their catalytic rates directly and strongly affect the growth of the organism. Any drastic change in these would make changes related to other enzymes negligible in comparison.

Figure 1.1-b shows a comparison of the growth prediction for the model with estimated enzymes (purple), all-average model (dark blue), and partial-average (light blue) models

designed to account for the missing enzymes⁶. For a better comparison, we also reproduce the vETFL results in orange on the same graph. An important feature to observe in the figure is the maximal uptake rate, or the rightmost point to each curve. Depending on the model, this point is higher or lower on the glucose uptake axis. The partial-average model (light blue) shows a higher predicted maximal glucose uptake than all the other models. Conversely, the all-average model (dark blue) shows a lower predicted maximal glucose uptake than all the other models. The original vETFL model (orange), and the vETFL model with estimated enzymes (purple) show a stopping point in between these two models with less information. This implies that enzymes with an influence on the maximal glucose uptake are accurately accounted for in the original vETFL model. Additionally, since the vETFL model (orange) has a lower maximal glucose uptake than an equivalent model with average enzyme concentrations (the partial-average model, light blue), then we can deduce that some limiting enzymes in the glucose metabolism have a k_{cat} parameter lower than the average value of $k_{\text{cat}}^{\pm} = 172 \text{ s}^{-1}$. Similarly, the full-partial model (dark blue) being more limited than the vETFL model (orange) hints that more enzyme information could reduce further the maximal glucose uptake the model can allow. The vETFL model with estimated enzymes (purple) shows an earlier plateau than vETFL (orange), its less constrained counterpart. Finally, we observe that the differences between these four models only appear at glucose uptake rates higher than $\approx 9 \text{ mmol}_{\text{glc}}.\text{DW}^{-1}.\text{h}^{-1}$, when the problem switches from being stoichiometry-limited to proteome-limited. Thus, this experiment illustrates the robustness of the formulation in predicting growth-limited phenotypes, but also the importance of well-curated catalytic rate constants for modeling organisms grown in proteome-limited regimens.

These results demonstrate the capability of ETFL to predict different phenotypes depending on growth rate. ETFL is also amenable to hypothesis testing, as evidenced using the models that estimate the missing enzymes. In particular, we showed with ETFL that an uptake increase does not yield a proportional growth rate increase as with FBA and that ETFL provides a maximal uptake rate that is unmodeled in FBA, thus more effectively modeling growth-dependent biomass yield in *E. coli*. This allows for more realistic predictions for phenotypes that are limited by the expression capabilities of the cell as well as captures the variability of the biomass composition in different growth regimens.

Variability analysis It is also possible to subject the model to a range of variability analyses. These are routinely used in FBA to assess the flexibility of the system and in TFA to find the ranges of allowed metabolite concentrations. In particular, we studied the number of bidirectional reactions in the system. Bidirectional reactions are reactions whose net flux can be either positive or negative. They are an indicator of the flexibility of the system. One of the main results of TFA was to replace ad-hoc assumptions on

⁶The following analysis has been adapted from the original text to match the current models and figure.

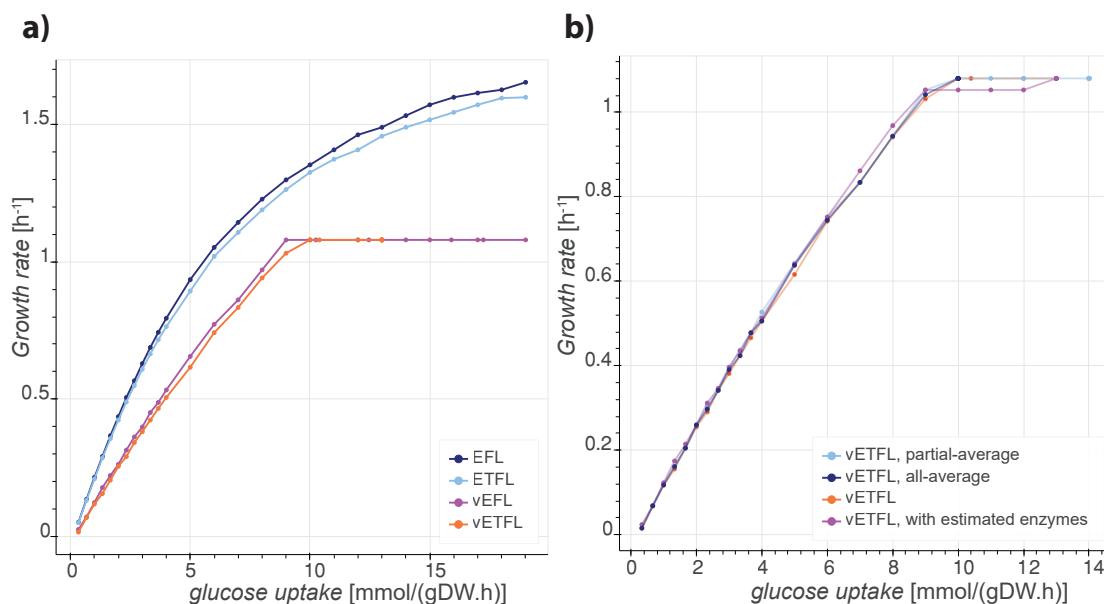


Fig. 1.1. Growth rate with respect to glucose uptake for differently constrained models in the ETFL framework. Legend in the same order as the height of the right-most point of each curve in each figure. **a.** Growth rate predictions using the EFL, ETFL, vEFL, vETFL models (dark blue, light blue, purple, orange); **b.** Growth rate predictions accounting for missing enzymes using vETFL (orange) and models (i)-(iii) (purple, light blue, dark blue) representing different initial enzyme assumptions, with k_{cat} values obtained from vETFL or $k_{\text{cat}} = 172 \text{ s}^{-1}$, and with/without inferred enzymes. Lines have transparency to better see overlaps.

the directionality of the reactions by thermodynamically-based directionality. We show that adding enzymatic constraints with ETFL also reduces the number of bidirectional reactions. The initial iJO1366 formulation with ad-hoc directionality assumptions shows 112 bidirectional reactions in FBA, under the constraint of a specific growth rate of 0.79 h^{-1} (TFA prediction). Once TVA is performed on the thermodynamics-enabled model of iJO1366, the number of bidirectional reactions drops to 88. Finally, after the addition of catalytic constraints, this number is reduced to 49 in the vETFL model.

We can extend the use of variability analyses in ETFL to explore the allowed proteome and transcriptome. For example, we measured the admissible extreme concentrations of each mRNA in aerobic growth conditions as described in McCloskey *et al.* (49) by performing a variability analysis on the mRNA concentration variables. Fig. 1.2 depicts the admissible peptide concentration upper and lower bounds, sorted by average, for vETFL with a glucose uptake set to $12.5 \text{ mmol.gDW}^{-1}.\text{h}^{-1}$, which yields a proteome-limited phenotype, according to our results in Fig 1.1a. It is important to note that all peptides with a non-zero minimal concentration (most of the left of the figure) are, by definition, essential peptides: These are always present at this uptake rate and are hence necessary for the cell to grow at an optimum growth rate. The same study can be performed for enzyme concentrations or even metabolite log-concentrations for models with thermodynamics. This type of study is useful for comparing how the model performs in relation to actual proteomics, transcriptomics, or metabolomics data. The method for

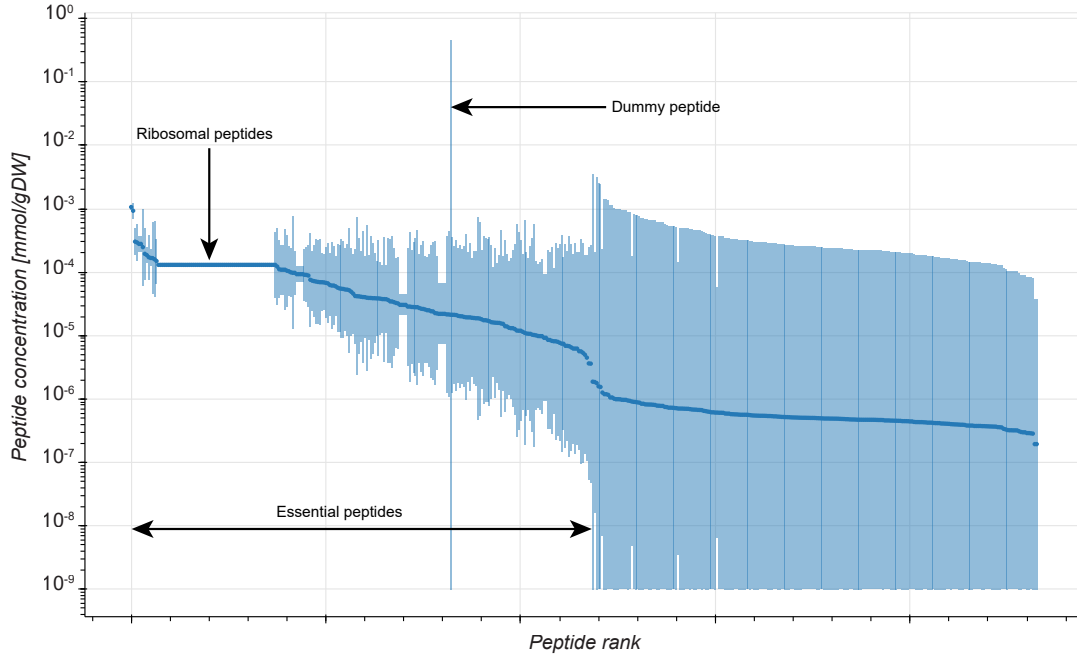


Fig. 1.2. Concentration variability of peptide species, sorted by average peptide concentration (darker disc). Lower bounds that were 0 were set to the accuracy of the solver, 10^{-9} . The horizontal line on the left side of the figure represents ribosomal peptides, which is narrow due to their instrumental role in making the tightly constrained amount of protein in the cell at a given growth rate. The vertical line in the middle represents the dummy peptide, which accounts for unmodeled peptides (non-metabolic proteins and enzymes with missing information) and therefore is used by the solver as a slack.

running these other types of variability analyses is exactly the same – only the variables subject to the variability analysis are changed.

A specific usage of a variability analysis is the study of the allowed proteome (resp. transcriptome) that is done by performing a variability analysis on the enzyme (mRNA) concentration variables. This type of study can, for instance, be compared with transcriptomics to check if the expression profile of an engineered strain corresponds to what is expected in its corresponding model. A way to visualize the average allowed proteome (transcriptome) is to use the average value of the variability of each enzyme (mRNA) concentration as a feasible observation⁷. This observation is then plotted on a finite area, which can be done using the online software Proteomaps(50, 51). This method and software are often used by biologists to represent protein abundances in the cell, and using the data from ETFL, we can generate similar comparative graphs that can help biologists

⁷Although there are no guarantees this point is part of the solution space, it remains a reasonable approximation of the average cell phenotype. If needed, we can verify it satisfies the constraints of the problem by imposing the concentration constraints and solving the obtained model. In case of infeasibility, a relaxation problem can easily be formulated to find the ℓ_1 -closest feasible point, in an approach similar to the *MOXA* methods presented in the section Adaptation of FBA-based methods to ETFL

analyze the variability in the different concentration variables using a visualization they are familiar with.

Fig. 1.3 is an example of such a representation, graphed using the mRNA concentrations corresponding to the solution represented by the dark dots in Fig 1.2 as an input. In this figure, mRNAs are clustered using KEGG Gene Ontology (GO) annotations. GO annotations form a tree describing the physiological role of genes, ranging from the least specific (e.g. general metabolism) to most specific (e.g. *araH* gene). The area of each (sub)cluster is proportional to the relative abundance of each (sub)group of mRNAs.

We used the mean of the variability analysis as the observation rather than a single optimal solution because the optimality principle in LP only guarantees a unique global optimum value and not a unique optimal solution. Moreover, solver heuristics give sparse and extreme results (corners of the explored simplex), which do not accurately represent the full extent of the considered solution space.

Essentiality analysis The ETFL framework can also analyze the essentiality of specific genes by performing single gene knockouts. The growth of models with knocked-out genes can then be compared to experimental data to assess the quality of the model as a validation⁸.

We performed a gene essentiality analysis using in ETFL and compared it to the results reported in the publication of iJO1366 by Orth *et al.* (48). We use the Matthew’s correlation coefficient (MCC) as a metric for the quality of the prediction, which is preferred over accuracy as it is not sensitive to the imbalance between the number of essential genes and non-essential genes. The MCC reads like a usual correlation coefficient, with 1 being a perfect correlation, -1 perfect anti-correlation, and 0 no correlation. We used the essentiality data and conventions given in the supplementary material of Orth *et al.*(48), as explained in Fig. 1.4-a and Fig. 1.4-b. The results are presented in Fig. 1.4-c,1.4-d, and 1.4-e, respectively for vETFL alone, vETFL supplemented by gene-protein association rules, and vETFL with estimated enzymes.

The vETFL model contains 563 enzymes requiring 475 gene expressions. For these genes, ETFL can perform a gene knockout analysis by setting their transcription rate to 0. The genes that do not participate in enzyme synthesis cannot be represented in such a manner,

⁸This section and Fig. 1.4 have been adapted to reflect the state of the newest models, and as such differs from the text and figure published in the published article. A major change is that subfigures Fig. 1.4-c now shows the results limited to the genes that express a peptide that participates in the composition of an enzyme. For the other genes, since the peptides they express do not participate in enzyme compositions (mostly because of missing data), their knock-out by setting the translation rate to 0 would not impact growth and they would be classified as non-essential. Subfigure Fig. 1.4-d captures the predictions of the model when the GPR is used for these genes that do not participate in the composition of any enzyme. Subfigure Fig. 1.4-e captures the predictions of the model when missing enzymes are estimated using the method detailed in the previous section.

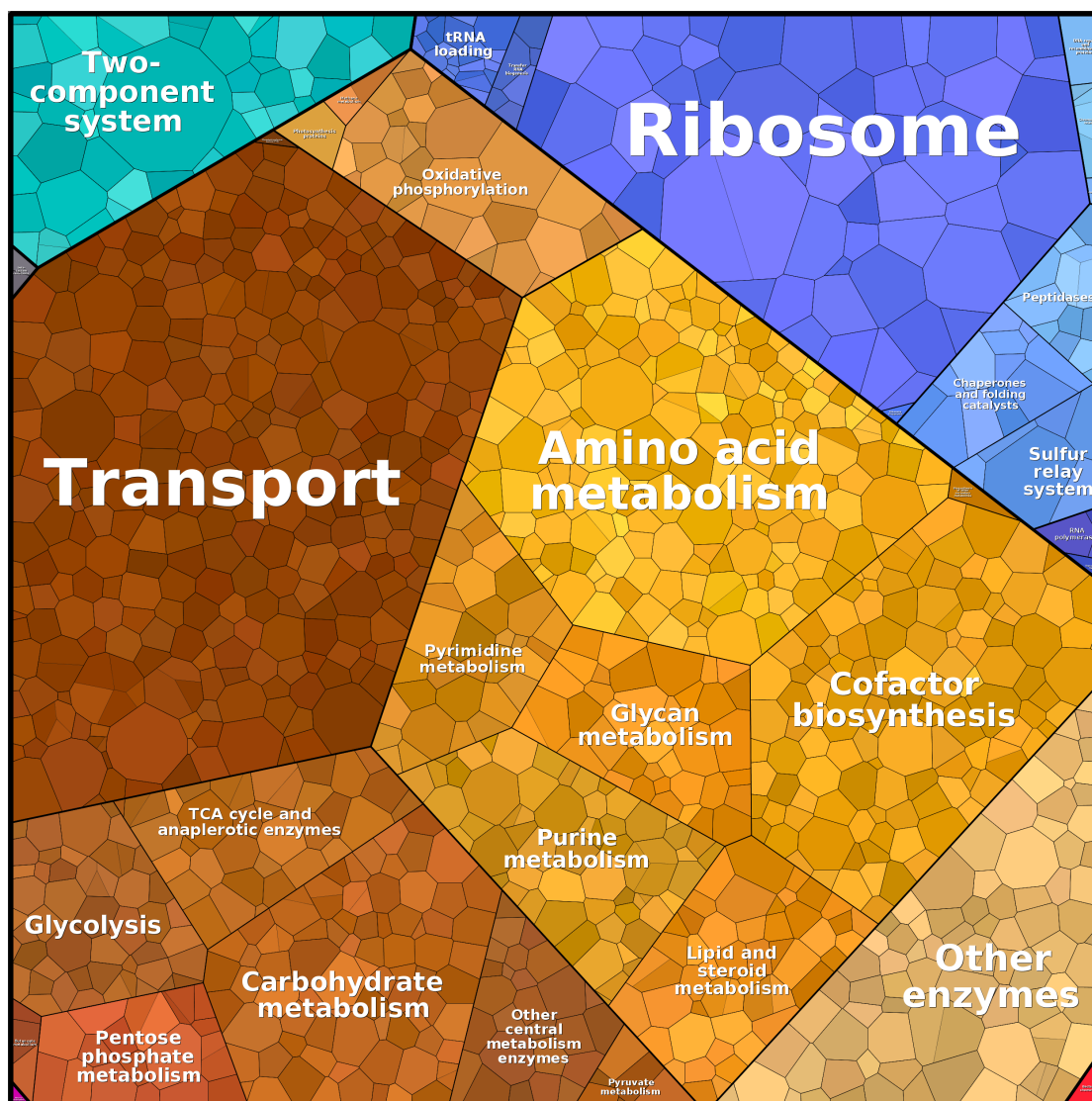


Fig. 1.3. Voronoi map of the predicted abundances of mRNAs. Each colored patch represents a different mRNA, with its area in proportion to its relative abundance. Genes can be clustered using KEGG Gene Ontology (GO) terms. Colors indicate the clustering.

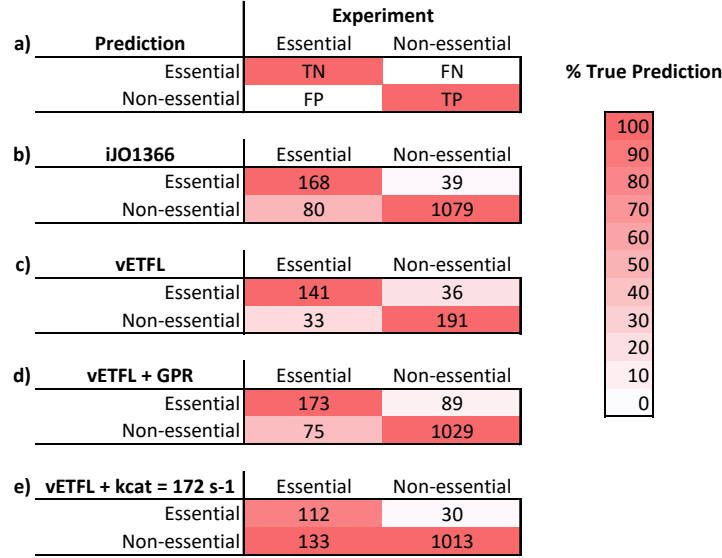


Fig. 1.4. Confusion matrices for gene essentiality studies. **a.** Conventions from Orth *et al.* (48) for gene essentiality. TN is True Negative. FN is False Negative. FP is False Positive. TP is True Positive. The color shading represents how good the classification is in the experimental class. Perfect classification should have a strict red first diagonal, as shown on this example. **b.** Gene essentiality prediction for the FBA model iJO1366, yielding a Matthew's correlation coefficient (MCC) of 0.69. **c.** Gene essentiality prediction for the genes expressed in the vETFL model, yielding a MCC of 0.65. **d.** Gene essentiality prediction for the vETFL model, where genes without enzyme assignment were tested using gene to protein to reaction (GPR) associations from the iJO1366 model, yielding a MCC of 0.61. **e.** Gene essentiality prediction for the vETFL model with estimated enzymes with all $k_{cat} = 172 \text{ h}^{-1}$, yielding a MCC of 0.54.

and are thus excluded from the analysis. On this reduced set of genes, ETFL performs essentially as well as iJO1366 (Fig. 1.4-b and 1.4-c), with a MCC of 0.65 (versus 0.69 for iJO1366), which indicates a good correlation between model results and experiments. Because of the strong difference in the size of the compared sets, direct comparison of particular elements of the confusion matrix is difficult. We thus produced two additional models, which account for the missing genes not represented in the expression problem of vETFL.

The first model uses gene-protein association rules (GPRs) to perform the knock-out on genes not represented in vETFL. GPRs are also what FBA-based essentiality analysis uses. With this model, we represent the essentiality predictions for 1366 genes (Fig. 1.4-d). The Matthew's correlation coefficient, at 0.61, is slightly lower than that of the previous model. Compared to iJO1366, we observe that this model predicts fewer false positives (experimentally essential genes predicted as non-essential), which means critical processes of the cell survival are better described. However, it also predicts more false negatives (experimentally non-essential genes predicted as essential). This means that redundant parts of the metabolism are lacking information and can be improved.

The second model was obtained by inferring enzymes from the GPRs, and setting their catalytic activities to an average value of 172 s^{-1} . We are able to represent the essentiality predictions of 1288 genes in this model (Fig. 1.4-e). The MCC yielded is worse (0.54) than that of the previous models, mostly because of the high false positive rate (genes predicted essential while being experimentally essential). However, the number of false negatives (genes predicted essential while being experimentally non-essential) is lower than in any other models. These results suggest that essential parts of the metabolisms are catalyzed by enzymes that differ from the average enzyme used for the inference, likely with a catalytic rate lower than 172 s^{-1} . Finally, changing the k_{cat} value used for enzyme inference will alter these numbers, with lower values (less efficient enzymes) increasing the number of predicted essential genes, and higher values increasing the number of non-essential genes. This can be understood as, at a given flux capacity, an enzyme with a lower catalytic rate constant will need to be more abundant, and thus mobilize more resources. Given the time needed to perform a gene essentiality analysis for a thousand of genes (in the order of a day), we did not venture to generate a ROC curve.

A detailed interpretation of the differences between gene knock-out in ETFL and FBA is discussed in Materials and Methods. The Supplementary Data provides more insights on the mismatched between ETFL essentiality results and iJO1366 essentiality results, and indeed shows that 87% of the mismatches are attributed to reactions without enzymatic data⁹. A significant fraction of mismatches (54%) come from the subsystems for the biosynthesis of lipids and cell envelope elements.

⁹These results are from the published article and were kept as such because they are still qualitatively valid

These results indicate that ETFL captures gene essentiality in well-characterized parts of metabolism. We also showed that an hybrid essentiality analysis using GPRs can be used without a significant degradation of the performance. Finally, inferring enzymes from GPRs is another way to account for genes without enzyme data, and the catalytic rate constant used for the inference is an important factor determining the essentiality of the gene.

Sampling Sampling the feasible solution space of FBA is a common way to study solution robustness and variability. Since there are often multiple FBA solutions at the optimal objective value, representative solutions are often sought, and sampling is one way to obtain them. However, because ETFL contains integer variables, it is not compatible with traditional sampling methods in its current formulation. It is possible, though, to make the model convex, and hence amenable to sampling, by fixing the integers to their values at a given growth rate and, if applicable, TFA directionality. This will block the flux directions as well as the growth-dependent parameters if TFA is performed. The resulting model is then solely linear, and sampling can be performed with traditional techniques, such as artificially centered hit and run (ACHR) (52), gpSampler (53), or optGpSampler (54). Once it has converged, a sampling should provide a better representation of the center of the solution space than the mean of the variability analysis.

1.2.3 Performance

For robustly reporting solution times of ETFL, we logged solving times each time a model was optimized during the redaction of this article. In that respect, some observations are the result of iterated optimizations, others from different optimization problems. In particular, variability and gene essentiality analyses require thousands of optimizations. We aggregated the solution times report the corresponding histograms, by model type, in Fig. 1.5. We measured the following metrics of the performance data: (i) arithmetic mean, (ii) geometric mean, and (iii) median. Although the distributions are not log-normal, it is common to report the geometric mean as a measure of the center of the distribution for comparison with other software (55, 56), as it is more robust to outliers than the arithmetic mean and more sensitive to unevenness than the median.

Using well-established MILP solvers (CPLEX (57), Gurobi (58)), we report a geometric mean solution time of 7.47 s for vETFL, with 95% of the problems solved in less than 100 s on the test hardware. This is 3 orders of magnitude better than the reported solution time for O'Brien *et al.* (40) (6 hours – 2×10^4 s) and between 1 and 2 orders of magnitude better than the reported solution time for Lloyd *et al.* using cobraME (41) (10 min – 6×10^2 s). It is worth noting that these vETFL optimizations also include thermodynamics constraints, which are absent of the other two formulations.

Chapter 1. The ETFL formulation allows multi-omics integration in thermodynamics-compliant metabolism and expression models

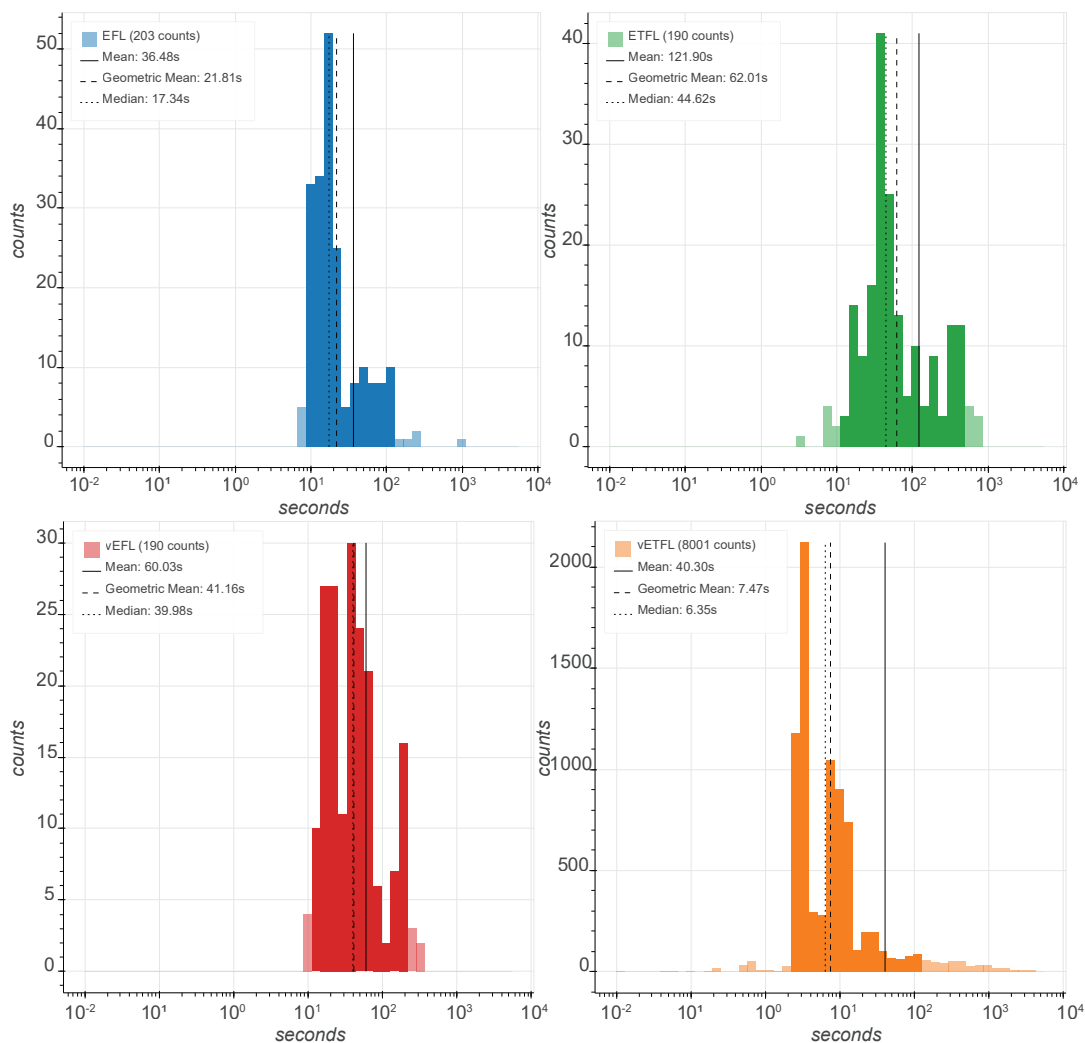


Fig. 1.5. Histograms displaying the distribution of solving times of each type of model during the data generation for this study. The darker area represents data between the 5th and 95th percentiles.

Table 1.4. Characteristic completion run times for several types of studies in the vETFL study of iJO1366

Study type (vETFL)	vETFL characteristic run time (h)
Growth curve (Fig. 1.1)	1
Enzyme VA	1.5
mRNA VA	2-3
Gene essentiality	24
50-points dETFL (see Adaptation of FBA-based methods to ETFL)	1

It is also important to state that although cobraME has an improved solution time over the original ME-model formulation, the formulation trades inequalities in the expression problem for equalities, and hence disregards a whole (non-growth optimal) part of the solution space that might contain physiological phenotypes. In particular, catalytic constraints become equalities, and the flux carried by reactions is set to be proportional to the amount of available enzyme instead of being upper-bounded by it. This gives less flexibility to the cell and prevents the representation of transient phenotypes. As an example, a cell that has been growing on a carbon source (*e.g.* glucose) will have a proteome suited to utilize this carbon source. However, once exhausted, it will need to reallocate its proteome to a new carbon source (*e.g.* lactose). In this transient state, some enzymes related to the first carbon source metabolism (*e.g.* glucose transporters) will carry no flux. In this case, cobraME would predict no flux, and also no enzyme concentration. In contrast, ETFL would allow for non-utilized enzymes and avoids such trade-offs, which is also crucial for accurately integrating proteomics data.

Such performance enhancements allow studies that would have been excessively time consuming using prior ME-model formulations. We show in Table 1.4 a list of typical completion times for common studies that require multiple optimizations to be carried out.

Finally, ETFL relies on solver-specific MILP algorithms and heuristics, which also means that great variability in performances can be observed depending on the solver parameters. We provide tuned presets for different tasks (gene knock-out, variability analysis, growth maximization) with the package, and recommend that users run their own solver tuning if long run times are observed. We witnessed an up to 10× increase in performance using such tuning.

1.2.4 Adaptation of FBA-based methods to ETFL

The ETFL formulation is amenable to further kinds of analyses. Leveraging both the explicit expression constraints and the MILP nature of the problem, we present several possibilities for future studies using ETFL:

Growth-dependent parameters It has been reported that several other parameters, such as the ribosome transcription rate constant k_{rib} , are growth dependent (47). Although such dependency is not taken into account in the presented results, it is possible to account for this by (i) discretizing k_{rib} following the method used to discretize the mRNA and protein content of the cell, and (ii) using Petersen’s linearization scheme (see Materials and Methods) on the product $k_{\text{rib}} * E_{\text{rib}}$. Other parameters that can be transformed in this way include, but are not limited to, the RNAP transcription rate constant k_{trans} , free ribosomes, and the RNAP ratios ρ and π .

Omics integration Explicit mRNA and enzyme concentrations allow the direct integration of absolute or relative proteomics and transcriptomics by changing the bounds of the corresponding variables in the EP. An additional gauge constraint will be needed for relative data. Previous transcriptomic integration methods, such as REMI (59), iMAT (60), GIMME (61), or MINEA (62), can also be adequately reformulated for ETFL. Metabolomics can still be integrated using TFA (35, 36).

Minimization of adjustment In its original article, the hypothesis behind the Minimization of Metabolic Adjustment (MOMA) method is that the metabolic fluxes of an organism subject to a gene knock out show a minimal change compared to the metabolic fluxes of the wild-type organism (63). The underlying hypothesis is that the enzyme distribution and assignments remain the same except for the knocked-out gene. With ETFL, it is possible to directly compute a Minimization of Protein Adjustment (MOPA) by reformulating the objective function as a Minimization of Expression Adjustment (MOXA):

$$\min \sum_{j \in \mathcal{J}} \|E_j - E_j^0\|_p, \quad p \in \{0, 1\} \quad (\text{MOPA})$$

where $\|\cdot\|_p$ is either the Manhattan norm ($p = 1$, ℓ_1 -norm) or the Euclidean norm ($p = 2$, ℓ_2 -norm), which will require a MIQP solver. In the same fashion, it is also possible to formulate a (weighted) Minimization of mRNA Adjustment (MORA) or even a Minimization of eXpression Adjustment (MOXA) using the following formulations:

$$\min \sum_{l \in \mathcal{L}} \|F_l - F_l^0\|_p, \quad p \in \{0, 1\} \quad (\text{MORA})$$

$$\min \theta \cdot \sum_{j \in \mathcal{J}} \|E_j - E_j^0\|_p + (1 - \theta) \cdot \sum_{l \in \mathcal{L}} \|F_l - F_l^0\|_p, \quad p \in \{0, 1\}, \theta \in [0, 1] \quad (\text{MOXA})$$

Parsimonious analysis Parsimonious FBA (pFBA)(53) was developed to address the high fluxes of some of the solutions given by FBA. Although this concern is addressed in

ETFL by the combined actions of the EP and thermodynamics, pFBA can be adapted to ETFL to study an organism under parsimonious constraints. For example, it is possible to reformulate it into a parsimonious expression problem to find the minimal expression level required to meet a growth target using objective functions similar to MOPA, MORA, and MOXA. It is also possible to turn the problem around to consider the allowed enzyme amounts under minimal flux constraint obtained by pFBA to assess the metabolic flexibility of an organism.

Dynamic ETFL (dETFL) Dynamic FBA (dFBA) (64) is a method that uses FBA to predict the dynamics of a biological system represented with a stoichiometric model. In its original static optimization approach (SOA) formulation, a FBA problem is solved at each time step. The value of boundary fluxes of the FBA problem are updated at each iteration with values produced with a kinetic law, such as Michaelis-Menten glucose uptake and oxygen diffusion. Because ETFL allows direct access to enzyme concentrations, it is possible to use the latter to reformulate dFBA in its SOA. The original SOA approach uses ad-hoc constraints on the absolute flux change at each time step. However, in ETFL, it is possible to bound flux changes indirectly by bounding enzyme and mRNA concentration changes in the EP. Effectively, this approach allows the movement from ad-hoc constraints to physiological constraints.

Use in kinetic frameworks Often, kinetic frameworks require a reference flux distribution as an input. ETFL can provide such a distribution, with an increased accuracy as compared to FBA.

1.2.5 Building an ETFL ME-model for other organisms

Building an ETFL model from a genome-scale model follows a detailed procedure, for which a SOP is provided in the Supplementary Note 2. In this procedure, it is the quality of the input data that will determine the accuracy of the model. A well-curated, elementally balanced model is a critical prerequisite. Since ETFL is essentially adding constraints to the FBA problem, it is important as well to ensure the feasibility of the initial model.

In ETFL, and ME-models in general, catalytic constraints are what links the metabolism to the expression problem. Because of this, the accuracy of the ETFL reconstruction is also heavily dependent on the quality of the catalytic rate constants k_{cat}^j . Such information is not always easily accessible. Hence, we recommend to at least manually curate the catalytic rate constants of the key parts of metabolism, namely (i) ATP synthase, (ii) RNA polymerase, (iii) ribosome. We also advise to pay attention to the pathways of the main carbon source metabolism, as small catalytic rate constants can heavily throttle

the rest of the metabolism. For missing catalytic rate constants, a placeholder value can be used. O’Brien *et al.* (40) used $k_{\text{cat}}^j = 65 \text{ s}^{-1}$, which is close to the median of the values used in the present study. In our comparison with inferred enzymes, we used $k_{\text{cat}}^j = 172 \text{ s}^{-1}$, which is the arithmetic mean of the data we gathered.

Another key component for catalytically constraining the model is to have quality enzyme composition information. Indeed, marking an enzyme as a monomer instead of a dimer halves its synthesis cost. A good source for this information is MetaCyc (65), and literature. As explained in the previous paragraph, special attention should be given to the ATP synthase, the RNA polymerase, the ribosome, and the enzymes of the main carbon pathway. Macromolecule degradation rates are less critical and can be averaged. Growth-dependent protein, RNA, and DNA ratios drastically improve the quality of the model, as they allow to account for the expression activity that is related to non-metabolic proteins.

In the construction of a model for another organism, approximating parameters based on values from an *E. coli* model should be done with care. Similarly to gap filling and the use of template reactions, conserving parameters across close species is helpful; however, conserving parameters across a large phylogenetic distance is erroneous. An example is the ribosome translation rate, which can vary by one order of magnitude between *S. cerevisiae* and *E. coli*.

Finally, great care should be taken with respect to the units. Different conventions are used across sources. Parameters for which this has been observed include catalytic rate constants, molecular weights, and concentrations.

1.3 Conclusion

ETFL is a framework which implements expression and thermodynamic formalism using mainstream double-precision MILP solvers. This could not be previously accomplished using state-of-the-art ME-models, which use specialized quad-precision solvers and do not support integer variables. The formalism itself is based on the explicit and direct relationship with the underlying biochemistry and provides a way to incorporate growth-dependent variables using MILP linearization techniques. These new growth-dependent variables provide a finer modeling of expression because they consider phenotypic differences in different growth regimens, which is key for accurate modeling. ETFL can also compute explicit mRNA and enzyme concentrations as well as perform direct -omics data integration. In this, ETFL complements and extends FBA capabilities by using explicit relationships in lieu of the typical assumptions on the relationships between the transcriptome, proteome, and fluxome. This explicit accounting of expression mechanisms provides a finer level of control and a more relevant prediction of gene-editing outcomes. ETFL is robust to missing data, as missing enzymes and their composition

can be approximated using average enzyme characteristics. Because of this and its operational similarity with classic FBA-related analyses, ETFL can be efficiently integrated in standard model-based pipelines. For this intent, we provide in the Supplementary Note 2 a standardized procedure to produce ETFL models from genome scale models. For example, metagenome-based genome-scale reconstructions such as published by Magnúsdóttir *et al.* (44) can be directly fed to the framework to generate models for each of the 773 bacteria they identified. Integration with platforms like KBase (66) can also be envisioned to automatically draft ETFL reconstructions parametrized by curated organism-specific data. In a more general way, ETFL can assess the allowed expression profiles of any biological system amenable to genome-scale modeling, such as in the metabolic engineering of biocatalysts, microbial communities, drug design, or personalized medicine.

1.4 Materials and Methods

1.4.1 Preliminaries, Conventions, and Notations

The mass balances of the macromolecules in ME-models is written with respect to their concentration variables. If we assume the cell is growing at a specific growth rate μ , we must assume that the volume of cell within which the mass balance is considered varies.

The mass balance of a macromolecule G will be written:

$$\frac{dm_G}{dt} = C_G \frac{dV_c}{dt} + V_c \frac{dC_G}{dt}, \quad (1.6)$$

$$= v_G^{\text{syn}} \cdot V_c - v_G^{\text{deg}} \cdot V_c, \quad (1.7)$$

where C_G is the concentration of the macromolecule C_G in the cellular volume V_c , for a total mass m_G in the cell, produced at a rate v_G^{syn} and degraded at a rate v_G^{deg} .

We next combine equations 1.6 and 1.7 and divide by V_c (necessarily non-zero) to write the time derivative of the concentration C_G :

$$\frac{dC_G}{dt} = v_G^{\text{syn}} - v_G^{\text{deg}} - \frac{1}{V_c} \frac{dV_c}{dt} \cdot C_G. \quad (1.8)$$

By definition, $\frac{1}{V_c} \frac{dV_c}{dt} = \mu$ is the specific growth rate of the cell (under the assumption of constant cell density ρ_c), and the term $\mu \cdot C_G$ is called the dilution term, or v_G^{dil} , as per Fredrickson's work on formulating growth models (67). It is a common assumption that the concentrations inside the cell remain time invariant (quasi-steady state assumption),

effectively yielding the constraint:

$$v_G^{\text{syn}} - v_G^{\text{deg}} - \frac{1}{V_c} \frac{dV_c}{dt} \cdot C_G = 0. \quad (1.9)$$

It is also understood from the formulation of the FBA that adding a new reaction to the system, such as:



results in adding terms to the mass balances of A and B :

$$\frac{d[A]}{dt} = \dots - \eta_A^j \cdot v_j, \quad (1.11)$$

$$\frac{d[B]}{dt} = \dots + \eta_B^j \cdot v_j. \quad (1.12)$$

The further extension of this to reactions of n reactants to m products is trivial.

Several parameter values are taken from the BioNumbers database (68). When used, we specify their identification number as well as the original source from which the value was reported. Finally, we will represent products between a parameter value and a variable by the symbol “ \cdot ” and products between two variables by the symbol “ $*$ ”.

Hereafter, we propose a detailed top-down approach to formulate the constraints being built for ETFL, starting from the metabolite network and moving down to RNA synthesis. The general organization for each macromolecule is to write down its mass balance, apply assumptions, and then detail its synthesis and consumption mechanisms.

Metabolites From FBA, the mass-balance relationship for metabolites can be written as:

$$S \cdot v = 0. \quad (\text{FBA})$$

For the rest of the formulation, it is necessary to split the net flux v from each reaction into its forward net component and backward net component:

$$v_j = v_j^+ - v_j^-, \quad v_j^+, v_j^- \geq 0. \quad (1.13)$$

Biochemical reactions are catalyzed by enzymes. Each enzyme (Enz_j) of concentration

E_j can catalyze a flux v_j subject to the enzyme capacity constraint, which is a function of its forward and backward catalytic rate constants $k_{\text{cat}}^{j,+}$ and $k_{\text{cat}}^{j,-}$:

$$0 \leq v_j^+ \leq k_{\text{cat}}^{j,+} E_j, \quad (1.14)$$

$$0 \leq v_j^- \leq k_{\text{cat}}^{j,-} E_j, \quad (1.15)$$

$$v_j^+ - k_{\text{cat}}^{j,+} E_j \leq 0, \quad (\text{FC}_j)$$

$$v_j^- - k_{\text{cat}}^{j,-} E_j \leq 0. \quad (\text{BC}_j)$$

The distinction between the bounds of the forward and backward net fluxes is important, as some enzymes have different catalytic activities depending on the direction of the flux.

1.4.2 General constraints for enzymes

Each enzyme Enz_j is represented by its total concentration, the variable E_j . It is subject to mass balance, which can be written:

$$\frac{d}{dt} E_j = v_j^{\text{asm}} - v_j^{\text{deg}} - v_j^{\text{dil}}, \quad (1.16)$$

which reads under quasi-steady state assumption (QSSA):

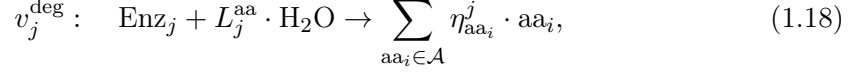
$$v_j^{\text{asm}} - v_j^{\text{deg}} - \mu * E_j = 0, \quad (\text{EB}_j)$$

where v_j^{asm} is the formation rate of the enzyme by the assembly of its constituent peptides, v_j^{deg} is the degradation rate, v_j^{dil} is the dilution rate, and μ is the growth rate of the cell. The formation rate of the enzyme describes the assembly of free peptides, hence it is necessary to add the peptide assembly reaction to the stoichiometric matrix:

$$v_j^{\text{asm}} : \quad \sum_{l \in \mathcal{L}} \eta_l^j \cdot \text{Pep}_l \rightarrow \text{Enz}_j, \quad (1.17)$$

where η_l^j is the stoichiometric coefficient of peptide Pep_l for the formation of the complex of enzyme Enz_j . This reaction is assumed to happen spontaneously by default.

We model the degradation reaction of the enzyme in the following manner:



where $\eta_{\text{aa}_i}^j$ is the number of aminoacids aa_i in the enzyme. It is obtained from the composition of the constituent peptides. For this degradation reaction, the rate is known:

$$v_j^{\text{deg}} - k_{\text{deg}}^j \cdot E_j = 0, \quad (\text{ED}_j)$$

where k_{deg}^j is the degradation rate constant of the enzyme. The reaction is added to the model and the equation ED_j is added as a constraint.

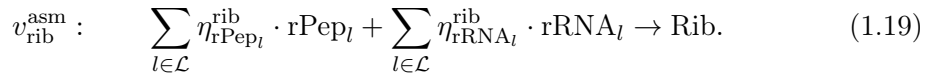
1.4.3 Constraints specific to Ribosomes

Like any other enzyme, ribosomes verify the mass balance:

$$v_{\text{rib}}^{\text{asm}} - v_{\text{rib}}^{\text{deg}} - \mu * E_{\text{rib}} = 0. \quad (EB_{\text{rib}})$$

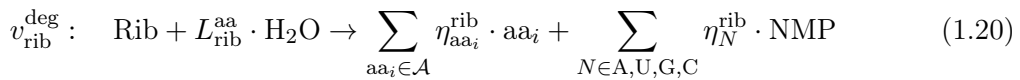
E_{rib} denotes the total quantity of ribosomes in a cell. It accounts for R_l , the ribosomes assigned to the translation of Pep_l , as well as the free ribosomes in the cell, R_{F} .

The ribosome differs from other enzymes in that it takes ribosomal peptides rPep_l as well as ribosomal RNA rRNA_l for its assembly. Hence, its assembly reaction is:



As explained earlier, the stoichiometric coefficients $\eta_{\star}^{\text{rib}}$ will appear in the mass balances of each of the compounds of the reaction. This reaction is also assumed to happen spontaneously by default

When ribosomes are degraded, their constituting amino acids and ribonucleotides are recovered:



The degradation rate is constrained in a manner similar to the constraint ED_j.

Finally, we can then write the total ribosome capacity constraint:

$$\sum_{l \in \mathcal{L}} R_l + R_F - E_{\text{rib}} = 0. \quad (\text{TC2})$$

If we know the ratio ρ of occupied vs free ribosomes, we can enforce it:

$$R_F - (1 - \rho) E_{\text{rib}} = 0. \quad (\text{RR})$$

1.4.4 Constraints specific to RNA Polymerase

RNAP is an enzyme, and hence it also satisfies mass balance:

$$v_{\text{RNAP}}^{\text{asm}} - v_{\text{RNAP}}^{\text{deg}} - \mu * E_{\text{RNAP}} = 0, \quad (EB_{\text{RNAP}})$$

where E_{RNAP} is the total amount of RNAP, which also accounts for free RNAP P_F . Its synthesis and degradation follow equations similar to other enzymes:

$$v_{\text{RNAP}}^{\text{asm}} - v_{\text{RNAP}}^{\text{deg}} - \mu * E_{\text{RNAP}} = 0, \quad (EB_{\text{RNAP}})$$

with the same conventions as in Eq. EB_j. As for a generic enzyme, RNAP is assembled from free peptides, which adds the peptide assembly reaction to the stoichiometric matrix:

$$v_{\text{RNAP}}^{\text{asm}} : \quad \sum_{l \in \mathcal{L}} \eta_l^{\text{RNAP}} \cdot \text{Pep}_l \rightarrow \text{RNAP}, \quad (1.21)$$

again with the same conventions as in the section General constraints for enzymes. This reaction is also assumed to happen spontaneously by default. The degradation reaction is also modeled similarly, with the same conventions:

$$v_{\text{RNAP}}^{\text{deg}} : \quad \text{Enz}_j + L_j^{\text{aa}} \cdot \text{H}_2\text{O} \rightarrow \sum_{\text{aa}_i \in \mathcal{A}} \eta_{\text{aa}_i}^j \cdot \text{aa}_i. \quad (1.22)$$

The degradation rate is constrained in a manner similar to the constraint ED_j.

Additionally, the total capacity of RNAP follows a capacity constraint similar to that of

ribosomes:

$$\sum_{l \in \mathcal{L}} P_l + P_F - E_{\text{RNAP}} = 0. \quad (\text{TC1})$$

As we did with the ribosomes, if we know the ratio of occupied RNAP, π , we can enforce it:

$$P_F - (1 - \pi) E_{\text{RNAP}} = 0. \quad (\text{PR})$$

1.4.5 Constraints for Peptides

The peptide concentrations obey the mass-balance equation:

$$\frac{d}{dt} \text{Pep}_l = v_l^{\text{tsl}} - \sum_{j \in \mathcal{J}} \eta_l^j \cdot v_j^{\text{asm}} - v_l^{\text{deg}} - v_l^{\text{dil}}. \quad (1.23)$$

We assume in the current model that the protein assembly rates are much faster than dilution and degradation, and thus simplify this mass balance to:

$$\frac{d}{dt} \text{Pep}_l = v_l^{\text{tsl}} - \sum_{j \in \mathcal{J}} \eta_l^j \cdot v_j^{\text{asm}}, \quad (1.24)$$

which, under QSSA, can be written:

$$v_l^{\text{tsl}} - \sum_{j \in \mathcal{J}} \eta_l^j \cdot v_j^{\text{asm}} = 0. \quad (\text{PB}_1)$$

In this context, the peptides are treated just like regular metabolites in the system. This assumption in PB_1 can be relaxed without a loss of generality by introducing a dilution and a degradation term, thus introducing a bilinearity.

The synthesis of peptides consumes charged tRNAs, which are subsequently uncharged during the current peptide synthesis by a ribosome. The process consumes 2 GTP and

releases 2 GDP and 2 Pi per amino acid:

$$\begin{aligned}
v_l^{\text{tsl}} : \quad & \sum_{aa_i \in \mathcal{A}} \eta_{aa_i}^l \cdot \text{tRNA}_{aa_i}^{\text{charged}} + 2L_l^{\text{aa}} \cdot (\text{GTP} + \text{H}_2\text{O}) \\
& \rightarrow \text{Pep}_l + \sum_{aa_i \in \mathcal{A}} \eta_{aa_i}^l \cdot \text{tRNA}_{aa_i}^{\text{uncharged}} + 2L_l^{\text{aa}} \cdot (\text{GDP} + \text{Pi} + \text{H}^+),
\end{aligned} \tag{1.25}$$

where aa_i denotes the i^{th} amino acid, $\eta_{aa_i}^l$ its stoichiometric coefficient (count) in the sequence of Pep_l , $\text{tRNA}_{aa_i}^*$ the (un)charged tRNAs for each amino acid, and $L_l^{\text{aa}} = \sum_{aa_i \in \mathcal{A}} \eta_{aa_i}^l$ is the length of the amino acid sequence of Pep_l .

As explained in the section Preliminaries, Conventions, and Notations, this reaction adds a supplementary term in the mass balances of the metabolites (GTP, GDP, Pi, H_2O , H^+), the peptide, and the tRNAs (see Constraints specific to tRNAs for the latter). This term is what connects the expression requirements to the metabolic network defined in the FBA.

The peptides are the product of a translation reaction that is catalyzed by a ribosome. As we did with the catalytic constraints for general biochemistry reactions, we can apply the ribosome maximum catalytic rate as an upper bound to its translation rate v_l^{tsl} :

$$v_l^{\text{tsl}} - \frac{k_{\text{cat}}^{\text{rib}}}{L_l^{\text{aa}}} R_l \leq 0, \tag{TR2_l}$$

where $k_{\text{cat}}^{\text{rib}}$ is the maximum ribosomal translation rate constant ($10 - 12 \text{ aa.s}^{-1}$ for *E. coli*, BioNumbers ID [BNID] 100059 (69)), L_l^{aa} is the amino acid length of the peptide l , and R_l is the concentration (in mmol.gDW^{-1}) of ribosomes assigned to the translation of this peptide. This way, the ratio R_l/Pep_l is effectively the number of ribosomes, or average polysome size, translating the peptide l .

1.4.6 Constraints for mRNAs

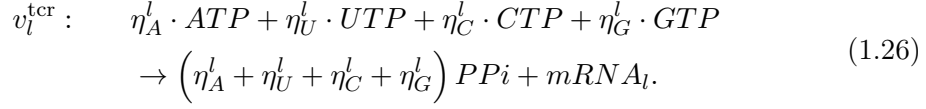
During the translation, an mRNA is read to produce a peptide. mRNAs are subject the same mass-balance constraints:

$$v_l^{\text{tcr}} - v_l^{\text{deg}} - \mu * F_l = 0, \tag{MB_l}$$

where F_l is the total concentration of the l^{th} mRNA ($mRNA_l$), v_l^{deg} is its degradation rate, and v_l^{tcr} is its transcription (synthesis) rate. F_l is variable that represents the

Chapter 1. The ETFL formulation allows multi-omics integration in thermodynamics-compliant metabolism and expression models

concentration of ($mRNA_l$). The transcription reaction is modeled as follows:



Again, the stoichiometric coefficients will appear in the mass balances of each of the metabolites and macromolecules involved. The transcription process is catalyzed by RNA polymerase (RNAP). For each transcription of mRNA, we can put an upper bound on the transcription rate v_l^{tcr} in the same way as for translation:

$$v_l^{\text{tcr}} - \frac{k_{\text{cat}}^{\text{RNAP}}}{L_l^{\text{nt}}} P_l \leq 0, \quad (\text{TR1}_l)$$

where L_l^{nt} is the length in nucleotides of the mRNA sequence, $k_{\text{cat}}^{\text{RNAP}}$ is the catalytic rate constant of RNAP (85 nt.s⁻¹ for *E. coli*, BNID 100060 (69)), and P_l the concentration of RNAP assigned to the transcription of this mRNA.

We must also take into account the relationship between ribosome assignment and mRNA concentration. On each strand of mRNA_{*l*}, there can be only a finite number ρ_l of ribosomes translating at the same time. This number is given by the ratio of the footprint size of the ribosome $L_{\text{rib}}^{\text{nt}}$ and the length of the mRNA strand L_l^{nt} . This effectively yields the number of ribosomes that can be present at the same time on a given mRNA strand:

$$\rho_l = \frac{L_l^{\text{nt}}}{L_{\text{rib}}^{\text{nt}}}. \quad (1.27)$$

For *E. coli*, $L_{\text{rib}}^{\text{nt}}$ is approximately 20 nm (BNID 102320 (70), 100121 (71)), which amounts to approximately 60 base pairs (the length of a nucleotide is approximately 0.3 nm; BNID 103777 (72)). From there we can get the additional constraint:

$$R_l \leq \rho_l \cdot F_l, \quad (1.28)$$

$$R_l - \frac{L_l^{\text{nt}}}{L_{\text{rib}}^{\text{nt}}} F_l \leq 0. \quad (\text{EX}_l)$$

We consider the following degradation reaction for mRNAs:



And, again, we know the degradation rates:

$$v_l^{\text{deg}} - k_{\text{deg}}^l \cdot F_l = 0. \quad (\text{MD}_l)$$

1.4.7 Constraints specific to rRNAs

rRNAs are used in the ribosome assembly reaction. According to the definition of $v_{\text{rib}}^{\text{asm}}$ in the Constraints specific to Ribosomes section, their mass balance can be written:

$$\frac{d}{dt} [\text{rRNA}_l] = 0 = v_{\text{rRNA}_l}^{\text{tcr}} - v_{\text{rib}}^{\text{asm}} - v_{\text{rRNA}_l}^{\text{deg}} - v_{\text{rRNA}_l}^{\text{dil}}. \quad (1.30)$$

We neglect their dilution and degradation under the hypothesis that free rRNAs are scarce and stable (73). Thus, their mass balance in the model reads:

$$v_{\text{rRNA}_l}^{\text{tcr}} - v_{\text{rib}}^{\text{asm}} = 0. \quad (\text{RB}_{\text{rRNA}_l})$$

The degradation reaction is the same as for mRNA, and is part of the total degradation of the ribosome.

1.4.8 Constraints specific to tRNAs

Since tRNAs are relatively stable molecules (73), we neglect their degradation. Let $T_{\text{aa}_i}^u$ (resp. $T_{\text{aa}_i}^c$) represent $[\text{tRNA}_{\text{aa}_i}^{\text{uncharged}}]$ (resp. $[\text{tRNA}_{\text{aa}_i}^{\text{charged}}]$). Then, we can write the following constraints:

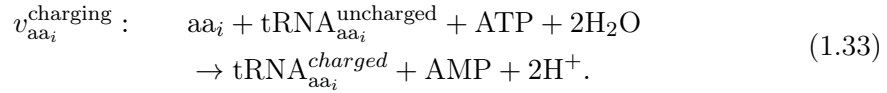
$$\frac{d}{dt}T_{aa_i}^u = 0 = -v_{aa_i}^{\text{charging}} + \sum_{l \in \mathcal{L}} \eta_{aa_i}^l \cdot v_l^{\text{tsl}} - \mu * T_{aa_i}^u, \quad (1.31)$$

$$\frac{d}{dt}T_{aa_i}^c = 0 = v_{aa_i}^{\text{charging}} - \sum_{l \in \mathcal{L}} \eta_{aa_i}^l \cdot v_l^{\text{tsl}} - \mu * T_{aa_i}^c, \quad (1.32)$$

$$-v_{aa_i}^{\text{charging}} + \sum_{l \in \mathcal{L}} \eta_{aa_i}^l \cdot v_l^{\text{tsl}} - \mu * T_{aa_i}^u = 0, \quad (\text{TB}_{aa_i}^u)$$

$$v_{aa_i}^{\text{charging}} - \sum_{l \in \mathcal{L}} \eta_{aa_i}^l \cdot v_l^{\text{tsl}} - \mu * T_{aa_i}^c = 0. \quad (\text{TB}_{aa_i}^c)$$

tRNAs are produced with a charging reaction and consumed by peptide synthesis. We use the following charging reaction:



By default, this reaction is assumed to happen spontaneously, but catalytic constraints can be applied if the adequate catalytic rate constants and enzyme compositions are known. Once again, the stoichiometric coefficients of each reactant will appear in the stoichiometric matrix in the column corresponding to this reaction.

1.4.9 Reformulation of the bilinearity of the problem

The main issue with the EP formulation presented previously lies in the continuous bilinear terms that describe the dilution of the macromolecules, $G_\star \in \{E_j\} \cup \{F_l\} \cup \{\text{tRNA}_{aa_i}^{(\text{un})\text{charged}}\}$. We use \star as a placeholder for the indexing of G . Using previous notations for the synthesis, degradation, and growth rate:

$$v_\star^{\text{syn}} - v_\star^{\text{deg}} - \mu * G_\star = 0. \quad (1.34)$$

In this state, the dilution term is bilinear, and the formulation requires a bilinear solver or potentially a mixed-integer bilinear solver if thermodynamics are to be added. The original ME-model formulation has similar terms as we are presenting here (39, 40). As such, its recent adaptation in Lloyd *et al.* (41) uses the two-level iterative algorithm SolveME (45) that requires a dedicated non-linear solver. In this fashion, iterative approaches which try to sequentially improve a value of the growth are a way to deal with the bilinearity. We present instead a MILP approximation of the problem that makes it compatible

and solvable with mainstream MILP solvers in a single optimization formulation. We achieve this through the discretization and linearization of the bilinear products. This operation can be understood as locally approximating the bilinear problem by several linear subproblems and choosing the best approximation.

Using a MILP approximation rather than an iterative scheme has two clear advantages. First, it allows to simulate growth-dependent parameters (such as RNA/Protein mass ratios) with guarantees on convergence and global optimality directly inherited from the MILP nature of the problem. In the case of parameters that are monotonically-increasing or decreasing with respect to growth rate, guarantees exist, such as showed in SteadyCom (74). However, in the case of non-monotonically increasing or decreasing parameters with respect to the growth rate, such guarantees are harder to prove, and thus MILP provides a strong framework to explore them, with global optimality guarantees and enumeration of alternative solutions. Second, by displacing the solving complexity to the solver, it also allows us to rely on the latest advances in MILP solving, which is a very dynamic field, with new solver releases every 6 to 12 months.

1.4.10 Approximation of the growth rate

In ETFL, we approximate the growth rate μ in bilinear products with a piecewise-constant function $\hat{\mu}$ (0^{th} order approximation). A zeroth-order approximation is an approximation by a piecewise-constant function. If $\hat{\mu}$ is piecewise-constant, then the product $\hat{\mu} * G_*$ is piecewise-linear. This can be represented in a MILP form, and allows us to transform the continuous bilinear terms into mixed (integer \times continuous) bilinear terms. This simplifies the problem, as these mixed bilinear terms can be linearized in a MILP setting using the Petersen linearization scheme (75), a particular case of the Glover linearization scheme (76) that was previously used in metabolic engineering by Hatzimanikatis *et al.* (77, 78).

Let $\bar{\mu}$ be an upper bound to μ , $(p, N) \in \mathbb{N}^2, p \leq N$. We can approximate μ with the following 0^{th} order approximation:

$$\forall \mu \in [0, \bar{\mu}], \quad \mu \approx \hat{\mu} = p \cdot \frac{\bar{\mu}}{N}. \quad (1.35)$$

With this notation, $\frac{\bar{\mu}}{N}$ is, in fact, the resolution of the approximation. N is the number of bins in which μ has been discretized, and p allows to choose which bin is selected in the solution. For the linearization of the problem, we will need to express p using only

binary variables. To this effect, we can perform its binary expansion:

$$p = \sum_{s=0}^{\lceil \log_2 N \rceil} 2^s \cdot \delta_s, \quad (1.36)$$

where $\lceil \log_2 N \rceil$ denotes the smallest majoring integer to $\log_2 N$, and $\delta_s \in \{0, 1\}$ is s^{th} digits from the right of the binary notation of p .

The model needs two more constraint to ensure that $\mu \in [\hat{\mu} - \frac{p}{N}, \bar{\mu} + \frac{p}{N}]$ and that p does not exceed N , which would result in $\hat{\mu} > \bar{\mu}$:

$$0 \leq \sum_{s=0}^{\lceil \log_2 N \rceil} 2^s \cdot \delta_s \leq N \quad (1.37)$$

$$-\frac{p}{N} \leq \mu - \hat{\mu} \leq \frac{p}{N}. \quad (1.38)$$

As an example, let us consider modeling an organism whose growth rate does not exceed $\mu_{max} = 2.3 \text{ h}^{-1}$. To do this, we can set $\bar{\mu} = 2.5 \geq \mu_{max}$. Let us choose a resolution of 0.25 h^{-1} , which gives $N = 10$. Then, $\log_2 N \approx 3.32$, and $\lceil \log_2 N \rceil = 4$. A growth rate $\mu = 1.4$ will be approximated by:

$$\begin{aligned} \hat{\mu} &= 1.5 = 6 \cdot \frac{\bar{\mu}}{10}, \\ \hat{\mu} &= (\delta_0 \times 2^0 + \delta_1 \times 2^1 + \delta_2 \times 2^2 + \delta_3 \times 2^3 + \delta_4 \times 2^4) \cdot \frac{\bar{\mu}}{10}, \\ \hat{\mu} &= (0 \times 1 + 1 \times 2 + 1 \times 4 + 0 \times 8 + 0 \times 16) \cdot \frac{\bar{\mu}}{10}. \end{aligned}$$

The values of δ_s are obtained by the solver upon optimization. This example is illustrated in Fig. 1.6-a. To maximize the resolution of the model, and minimize the associated computational cost (under the form of 3 additional constraints for each linearization to be performed, see Petersen linearization in Methods), the user should ideally choose N as a power of 2.

MILP solvers use a variety of algorithms and heuristics to solve MILP problems. In this case, the difficulty lies in the fact that the EP and the FBA are almost independent and linked through a limited number of equations and variables. Even though the automated solving methods of the solver might seem obscure to a human, we thought useful to provide a human-understandable heuristic for solving a formulation such as ETFL. It

might prove useful in the case where one needs to find an initial non-optimal solution, which sometimes greatly improve solver performances. Thus, conceptually, a heuristic for solving an ETFL problem would be:

1. Solve the FBA for μ
2. Select the corresponding, closest $\hat{\mu}$
3. Apply it to compute dilution values
4. Solve the EP with fixed dilution
5. Apply the catalytic constraints to the FBA
6. Recalculate the FBA under catalytic constraints
7. If $\mu \notin \left\{ \hat{\mu} \pm \frac{\bar{\mu}}{N} \right\}$, go back to 3, else, end.

1.4.11 Linearizing the bilinearity

In the previous derivation, we replaced the growth rate variable by a discrete number of acceptable values. We can approximate the continuous product $\mu * G_\star$, which represents the dilution, as follows:

$$\mu * G_\star \approx \hat{\mu} * G_\star, \quad (1.39)$$

$$\hat{\mu} * G_\star = \sum_{r=0}^{\lceil \log_2 N \rceil} \frac{2^r}{N} \cdot \delta_s * G_\star, \quad (1.40)$$

The product $\delta_l * G_\star$ is then still bilinear, but one of its variables is binary. Assuming a constant $M > G_\star$, We can use Petersen's linearization theorem (75, 76) to replace the product $\delta_s * G_\star$ with a single nonnegative variable z_\star^s , as described in the section Petersen linearization.

Because of the binary expansion, the complexity of the model grows only as $\mathcal{O}(\log_2 N) = \mathcal{O}(\log_2 \frac{1}{\epsilon})$, where $\epsilon = 1/N$ is proportional to the resolution of the approximation (which is $\frac{\bar{\mu}}{N}$). This means that the linearization part of a model with a resolution of 0.01 h^{-1} is only around twofold bigger than that of a model with a resolution 0.04 h^{-1} , while resolution has been improved fourfold.

1.4.12 Petersen linearization

After discretization of the growth rate, the dilution term for the macromolecule G_\star will consist of a sum of products of the binary variables δ_s and the continuous variable G_\star .

Chapter 1. The ETFL formulation allows multi-omics integration in thermodynamics-compliant metabolism and expression models

We can use the Petersen linearization scheme (75) to transform this product into an equivalent system of one new variable and three new constraints:

$$\begin{aligned}
 z_\star^s &= \delta_s * G_\star, \\
 &\iff \begin{cases} G_\star + M \cdot \delta_s - M \leq z_\star^s \leq M \cdot \delta_s, \\ z_\star^s \leq G_\star \end{cases}, \\
 &\iff \begin{cases} G_\star + M \cdot \delta_s - z_\star^s \leq M, \\ z_\star^s - M \cdot \delta_s \leq 0, \\ z_\star^s - G_\star \leq 0. \end{cases}
 \end{aligned} \tag{1.41}$$

With this method, we can directly reformulate generalized mass balances as described in Eq. 1.34 for mRNAs, enzymes, uncharged tRNAs, and charged tRNAs:

$$v_j^{\text{asm}} - v_j^{\text{deg}} - \sum_{r=0}^{\lceil \log_2 N \rceil} \frac{2^s}{N} \bar{\mu} \cdot z_j^s = 0, \tag{EB'_j}$$

$$v_l^{\text{tcr}} - v_l^{\text{deg}} - \sum_{r=0}^{\lceil \log_2 N \rceil} \frac{2^s}{N} \bar{\mu} \cdot z_l^s = 0, \tag{MB'_1}$$

$$-v_{\text{aa}_i}^{\text{charging}} + \sum_{\text{aa}_i \in \mathcal{A}} \eta_{\text{aa}_i}^l \cdot v_l^{\text{tsl}} - \sum_{r=0}^{\lceil \log_2 N \rceil} \frac{2^s}{N} \bar{\mu} \cdot z_{\text{aa}_i}^{u,s} = 0, \tag{TB'_{\text{aa}_i}^u}$$

$$v_{\text{aa}_i}^{\text{charging}} - \sum_{\text{aa}_i \in \mathcal{A}} \eta_{\text{aa}_i}^l \cdot v_l^{\text{tsl}} - \sum_{r=0}^{\lceil \log_2 N \rceil} \frac{2^s}{N} \bar{\mu} \cdot z_{\text{aa}_i}^{c,s} = 0. \tag{TB'_{\text{aa}_i}^c}$$

And we get the additional linearization constraints:

$$\sum_{r=0}^{\lceil \log_2 N \rceil} \frac{2^s}{N} \bar{\mu} \leq \bar{\mu}, \tag{GR}$$

$$-\frac{\bar{\mu}}{2N} \leq \mu - \sum_{r=0}^{\lceil \log_2 N \rceil} \frac{2^s}{N} \bar{\mu} \leq \frac{\bar{\mu}}{2N}. \tag{GC}$$

1.4.13 Discretization of mRNA and enzyme content

Since growth has been discretized, it is now possible to also directly discretize other growth-dependent parameters of the problem, regardless of whether they are in a linear or non-linear relationship with growth. This is a direct consequence of the formulation of ETFL, which allows some flexibility in the modeling assumptions of the user. As an example, we described the relationship between growth and protein and mRNA mass ratios, P^m and R^m , in the cell as reported in Neidhardt *et al.* (47). This relationship is described in the formulation section as constraints IC_{Enz} and IC_{mRNA}. We thus aim to approximate the non-linear function $P^m(\mu)$ (resp. $R^m(\mu)$) over the interval $[0, \bar{\mu}]$ with a piecewise-constant function $\widehat{P^m}$ (resp. $\widehat{R^m}$). We perform this approximation by interpolating and discretizing the protein ratio and mRNA ratio as functions of the growth rate so that:

$$\widehat{P^m} = \sum_{u \in \mathcal{U}} \lambda_u \cdot P_u^m, \quad (1.42)$$

$$\widehat{R^m} = \sum_{u \in \mathcal{U}} \lambda_u \cdot R_u^m, \quad (1.43)$$

where $P_u^m = P^m(u \cdot p \frac{\bar{\mu}}{N})$ (resp. $R_u^m = R^m(u \cdot p \frac{\bar{\mu}}{N})$). λ_u are binary variables, and only one can be active at a time, since we are choosing exactly one value per function. To enforce this behavior, we used a special ordered set constraint of type 1 (SOS1):

$$\sum_{j \in \mathcal{J}} MW_j \cdot E_j - \sum_{u \in \mathcal{U}} \lambda_u \cdot P_u^m = 0, \quad (\text{IC1})$$

$$\sum_{l \in \mathcal{L}} MW_l \cdot F_l - \sum_{u \in \mathcal{U}} \lambda_u \cdot R_u^m = 0, \quad (\text{IC2})$$

$$\sum_{u \in \mathcal{U}} \lambda_u = 1. \quad (\text{SOS1})$$

P_u^m and R_u^m are growth-dependent, interpolated protein and RNA mass ratios (in $\text{g} \cdot \text{g}^{-1}$). Given a growth rate, they define the relative mass of the cell that is protein or RNA. MW_\star represents the molar weight of the corresponding enzyme or RNA, and this their product with macromolecules concentrations (in $\text{mmol} \cdot \text{gDW}^{-1}$) will result in mass ratios as well, in grams per gram of dry cell weight. The first two constraints enforce equality between the interpolated data and the model production. The last line is the SOS1 constraint that forces only one of the λ_u to be active.

Additionally, it is necessary to have the integer index of λ_u equal to the index of the

growth rate. This is obtained through the constraint:

$$\sum_{u \in \mathcal{U}} u \cdot \lambda_u - \sum_{l \in \mathcal{L}} 2^l \cdot \delta_l = 0. \quad (\text{EQI})$$

The first term represents the growth integer index (which discrete value of $\hat{\mu}$ to use for choosing P_u^m), and the second represents its binary expansion (which discrete value of $\hat{\mu}$ to use for μ). The constraint makes sure they are equal.

Imposing such mass ratios requires the addition of a dummy mRNA as well as a dummy protein to represent the part of the transcriptome/proteome that is either missing from the expression model or altogether unrelated to metabolic function. We use average amino acid frequencies and GC content to model this. Explicit interpolation functions can also be used, such as the growth-dependent functions given in Pramanik *et al.* (79).

The simultaneous use of catalytic constraints on metabolic reactions (Eq. FC_j, BC_j) and maximal enzyme load (Eq. IC1) effectively implements allocation constraints like in GECKO (38), although in ETFL, the enzyme concentrations are also directly linked to the metabolism. In GECKO, the metabolic cost of building the enzymes is not taken into account.

Fig. 1.6-b shows an example piecewise linear interpolation of the growth-dependent protein mass ratio in *E. coli* according to Neidhardt *et al.* (47). The reported values (red circles) are interpolated using a piecewise linear function (dashed line), which is then discretized (full line). Using the integer constraints described above, the model can be forced to display a protein content that corresponds to its growth. We apply the same techniques to mRNA and DNA content.

1.4.14 Discretization of DNA content

To further increase the scope of macromolecules covered by the model, it is also possible to add growth-dependent DNA content, as expressed in the constraint IC_{DNA} of the formulation. DNA mass ratios at specific growth rates are reported in Neidhardt *et al.* (47). We model the DNA reaction synthesis as follows:

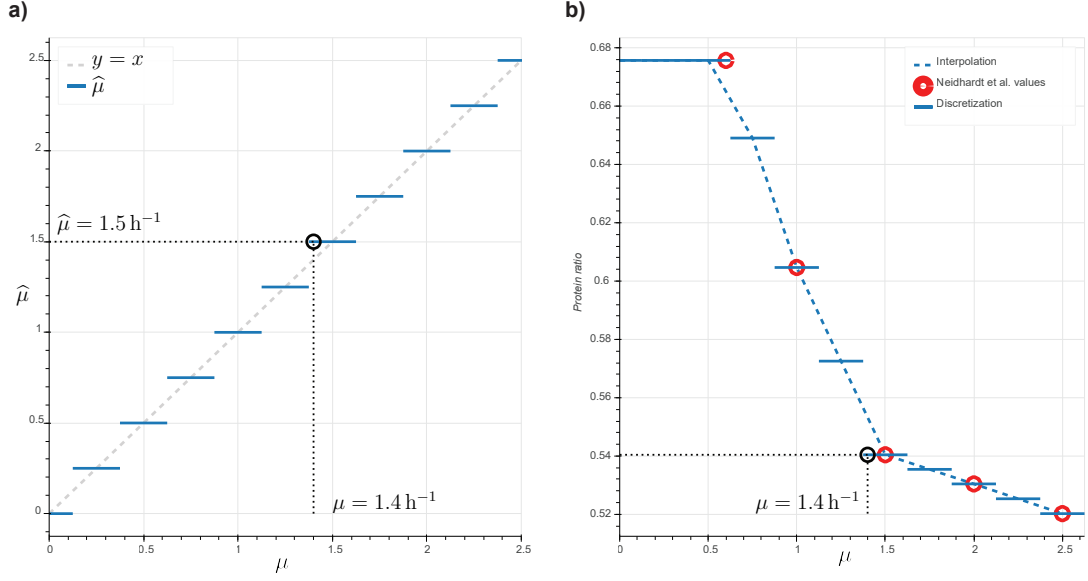


Fig. 1.6. Discretization example for specific growth rate and growth-dependent parameters. **a.** Discretization of μ into $\hat{\mu}$. The step approximation transforms the continuous interval $[0, 2.5]$ into the discrete set $\{0, 0.25, \dots, 2.5\}$. **b.** Example of piecewise linear interpolation and discretization of the protein mass ratio from Neidhardt *et al.* (47). Red circles represent the values reported. The dashed line is the piecewise linear interpolation. The solid line is its discretization.

$$\begin{aligned}
 v_{\text{DNA}}^{\text{synthesis}} : \quad & (1 - \gamma) L_{\text{DNA}}^{\text{bp}} d\text{ATP}, \\
 & + (1 - \gamma) L_{\text{DNA}}^{\text{bp}} d\text{TTP}, \\
 & + \gamma L_{\text{DNA}}^{\text{bp}} d\text{GTP}, \\
 & + \gamma L_{\text{DNA}}^{\text{bp}} d\text{CTP}, \\
 & \rightarrow \text{DNA} + 2L_{\text{DNA}}^{\text{bp}} \text{PPi},
 \end{aligned}$$

where γ is the GC content of the cell, and $L_{\text{DNA}}^{\text{bp}}$ is the total length in base pairs of the DNA. As with mRNA_l and Enz_j , DNA has a mass-balance equation of the following shape:

$$\begin{aligned}
 \frac{d}{dt} [\text{DNA}] = 0 &= v_{\text{DNA}}^{\text{synthesis}} - v_{\text{DNA}}^{\text{degradation}} - v_{\text{DNA}}^{\text{dilution}}, \quad (1.44) \\
 v_{\text{DNA}}^{\text{synthesis}} - v_{\text{DNA}}^{\text{deg}} - \mu * \text{DNA} &= 0. \quad (\text{DB}_{\text{DNA}})
 \end{aligned}$$

We consider that the DNA does not degrade, meaning the only source of DNA consumption is dilution caused by the growth of the cell and $k_{\text{deg}}^{\text{DNA}} = 0$. We then define the molar weight of DNA MW_{DNA} and enforce the DNA mass ratio Dm as we did with both proteins

and mRNA:

$$\begin{aligned} \text{MW}_{\text{DNA}} = (1 - \gamma) L_{\text{DNA}}^{\text{bp}} (\text{MW}_{\text{dATP}} + \text{MW}_{\text{dTTP}}), \\ + \gamma L_{\text{DNA}}^{\text{bp}} (\text{MW}_{\text{dGTP}} + \text{MW}_{\text{dCTP}}), \end{aligned} \quad (1.45)$$

$$\text{MW}_{\text{DNA}} \cdot \text{DNA} - \sum_{u \in \mathcal{U}} \lambda_u \cdot \text{Dm}_u = 0. \quad (\text{IC3})$$

If DNA is modeled, we can also include a catalytic constraint on its synthesis by DNA Polymerase¹⁰. We model the DNA Polymerase III holoenzyme according to the structure reported by Kelman *et al.* (80), as a dimeric enzyme attached by a τ scaffold to two β clamps sliding on the DNA. We assume it synthesizes DNA at a speed of 1 kilobase per second, according to the same source. The catalytic constraint on DNA synthesis is hence:

$$v_{\text{DNA}}^{\text{synthesis}} - \frac{k_{\text{pol}}^{\text{DNAPol3}}}{L_l^{\text{bp}}} \text{DNAPol3} \leq 0, \quad (\text{DP})$$

with $v_{\text{DNA}}^{\text{syn}}$ the synthesis rate of the DNA, $k_{\text{pol}}^{\text{DNAPol3}}$ the above-mentioned nucleotide synthesis rate, L_l^{bp} the length of the DNA in base pairs, and DNAPol3 the concentration of DNA Polymerase in the cell.

1.4.15 Gene copy number and RNAP saturation

In the same fashion as we considered the saturation of a mRNA strand by ribosomes, we can consider the saturation of gene open reading frames (ORFs) by RNA Polymerases¹¹. This is important to account for the phenomenon of plasmid burden and RNAP competition in recombinant organisms, as well as the effect of gene copy numbers on the transcription capacity of the cell (81, 82).

For *E. coli*, the footprint size of the RNA Polymerase $L_{\text{RNAP}}^{\text{nt}}$ is approximately 40 nucleotides wide (BNID 107873, (83)). Similarly to Eq. EX₁, we can write:

$$P_l \leq \frac{L_l^{\text{nt}}}{L_{\text{RNAP}}^{\text{nt}}} G_l \quad (1.46)$$

¹⁰This paragraph has been added in this thesis, to account for developments in the formulation post-publication.

¹¹This section has been added in this thesis, to account for developments in the formulation post-publication.

where we recall P_l is the concentration of polymerase allocated to the transcription of the l^{th} gene, L_l^{nt} the length of the ORF in nucleotides, and G_l the concentration of ORFs of the l^{th} gene. The latter is exactly equal to the concentration of DNA times the number of copies n_l of the gene:

$$G_l = n_l \cdot \text{DNA}. \quad (1.47)$$

We can thus derive a constraint for each gene:

$$P_l - \frac{L_l^{\text{nt}}}{L_{\text{RNAP}}^{\text{nt}}} \cdot n_l \cdot \text{DNA} \leq 0. \quad (\text{CN}_l)$$

1.4.16 Expression constraints for genes without enzymes

The vETFL model has 475 genes out of 1433 that participate in the composition of enzymes¹². The genes that do not participate in the composition of enzymes in the model can either (i) still be used to generate transcription and translation reactions, mRNAs, peptides, and the related constraints, or (ii) be ignored and solely used for their gene-protein association rule. Option (i) is preferable if transcriptomics or proteomics data are available, as the presence of the variable and constraints related to these genes will improve the quality of the 'omics integration. Option (ii) is preferable in the absence of transcriptomics and proteomics data, as it will reduce the number of equations in the problem. Indeed, each gene that is expressed in the model is linked to the constraints MB_l, PB_l, MD_l, TR1_l, TR2_l, EX_l, CN_l, the associated linearization constraints (3 per macromolecule), and the associated variables. Thus, including genes without a corresponding enzyme and without corresponding 'omics contributes significantly to the growth of the problem, without increasing the accuracy of the model.

1.4.17 Scaling

A critical issue in the formulation of this problem is that the variables are different orders of magnitude. Fluxes are typically between $10^{-3} - 10^1 \text{ mmol.gDW}^{-1}.\text{h}^{-1}$, whereas protein concentrations are around $10^{-6} - 10^{-3} \text{ mmol.gDW}^{-1}$ and mRNA concentrations are $10^{-10} - 10^{-6} \text{ mmol.gDW}^{-1}$. The relationship between these scales is given by the catalytic rate constant of enzymes and expression machinery, which spans from $10^3 - 10^6 \text{ h}^{-1}$. In particular, the ribosome rate constant for translation ($\sim 12 \text{ aa.s}^{-1} = 43\,200 \text{ aa.h}^{-1}$) as well as the RNA polymerase rate constant of transcription ($\sim 85 \text{ nt.s}^{-1} = 306\,000 \text{ nt.h}^{-1}$) are

¹²This section has been added in this thesis, to account for developments in the formulation post-publication. The remark applies to the vETFL model used in the graphs of this these, not the paper's vETFL model

responsible for strong differences in the concentrations and fluxes between transcription- and translation-related parts of the problem. Consequently, the constraint matrix becomes ill-conditioned, and the solver has to operate close to, or sometimes beyond, its maximal solving accuracy (usually around 10^{-9} for commercial solvers such as ILOG CPLEX or Gurobi).

To circumvent these limitations, we scale the EP, which will reduce the numerical difficulty of the problem, using nondimensionalization. We create nondimensionalized variables by dividing the variables of the initial problem by an estimated upper bound. For example, by definition, macromolecule concentrations cannot exceed 1 g.gDW^{-1} , and the following constrains the transformed macromolecule variables between 0 and 1:

$$\hat{X} = \frac{X}{\sigma_X}, \sigma_X \geq \sup(X) \implies 0 \leq \hat{X} \leq 1. \quad (1.48)$$

In this scheme, σ_X is an upper bound to X . In particular, if we consider σ_X to be the concentration of 1 g.gDW^{-1} :

$$\begin{aligned} \sigma_X &= 1 \text{ g.gDW}^{-1}, \\ &= 1 \text{ g.gDW}^{-1} \times \frac{1}{\text{MW}(X)} \text{ mmol.g}^{-1}, \\ &= \frac{1}{\text{MW}(X)} \text{ mmol.gDW}^{-1}, \end{aligned} \quad (1.49)$$

where $\text{MW}(X)$ denotes the molecular weight of the macromolecule in SI units ($\text{kg.mol}^{-1} \equiv \text{g.mmol}^{-1}$), and \hat{X} represents the mass fraction of the molecule in the cell. We scale the fluxes using a method derived from this, detailed in the supporting file Supplementary Note 1. It is also possible to further refine this upper bound by performing a variation analysis on X and re-generating a model using the newly estimated upper bound.

For the sake of clarity, all problem formulations will be kept in their dimensionalized form in the subsequent equations although the implementation is in fact nondimensionalized. The nondimensionalized problem is described further in Supplementary Note 1.

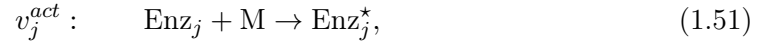
1.4.18 Advanced modeling

ETFL is amenable to modeling more intricate expression processes. A short selection of these is detailed below.

Enzyme-mediated complex assembly By default, all the peptides are assumed to assemble spontaneously, without an enzyme. However, in the case of an enzyme-mediated assembly, it is possible to limit the assembly rate by a catalytic constraint if needed, in a fashion similar to Eq. 1.14. If we denote A the total concentration of assembling enzyme, and k_{asm}^A the catalytic rate constant of assembly, we can constraint v_j^{asm} the assembly rate of the j^{th} enzyme:

$$v_j^{asm} \leq k_{asm}^A \cdot A. \quad (1.50)$$

Enzyme activation and post-translational modifications Some enzymes require to be modified in order to be active, and sometimes by metabolites of the cell. This can be captured by adding a new species representing the active enzyme, and an activation reaction transforming the inactive enzyme to the active form. If the metabolite M is required to activate enzyme Enz_j into Enz_j^* , then the following activation reaction is added to the model:



The mass balances of Enz_j and M will be supplemented by a term $-v_j^{act}$, and the mass balance of Enz_j^* by $+v_j^{act}$. Finally, the catalytic constraint of the reaction v_j catalyzed by Enz_j^* at concentration E_j^* shall be:

$$v_j \leq k_{cat}^j \cdot E_j^* \quad (1.52)$$

This reaction can be catalytically limited if needed (see previous paragraph), and require the participation of metabolites. Thus, ETFL allows to capture protein-metabolite interactions.

Enzyme association It is also possible to model the partition between free enzymes and associated enzymes. In that case, we simply need to operate the following adaptations: (i) replace the E_j term in any catalytic constraint by a new variable E_j^r , which represents the enzymes participating in the catalysis of the j^{th} reaction; (ii) add a variable E_j^F which represents the free enzymes of the system; and (iii) add the enzyme usage constraint:

$$E_j^r + E_j^F - E_j = 0 \quad (\text{EU}_j)$$

Dilution and degradation assumptions In the current formulation, some species have their dilution or degradation neglected because of high reactivity or slow degradation rate constants. This can be relaxed by simply editing the mass balance reaction according to the assumption to be relaxed. In particular, enzyme-mediated degradation can be modeled by adding suitable catalytic constraints on the degradation reactions. Additionally, the dilution term for metabolites can be taken into account if needed, in a manner similar to what Benyamini *et al.* describe in their method for FBA accounting for dilution (84).

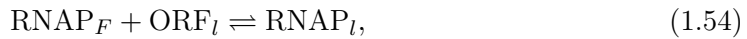
MILP-based gene knock-out strategies for strain design The ETFL formulation of gene knock-out using an upper bound on the translation rate allows to directly formulate MILP-based gene knock-out strategies for strain design. Indeed, for each l^{th} gene, we can enforce the constraint:

$$v_l^{\text{tsl}} \leq M \cdot b_l, \quad (1.53)$$

with v_l^{tsl} the gene's transcription rate, b_l a binary variable and M a big-M constant. With that kind of constraint, if $b = 1$, the gene is active, while if $b = 0$, the gene is knocked-out. It is hence possible to formulate an objective function to optimize the number of KO while fulfilling a metabolic objective, for instance.

Thermodynamic equilibrium of RNAP with promoters The binding of RNAP to promoters is an event that follows thermodynamic equilibrium laws (85), and has been successfully modeled before (81)¹³. In particular, it is possible to take advantage of the discretization method to model this equilibrium in ETFL.

Given the following binding reaction:



where RNAP_F is the free RNAP, RNAP_l the RNAP bound to ORF_l , the open reading frame of the l^{th} gene. We can write the binding constant K_B (86) of this reaction using the previous notations for RNAP concentrations, and G_l as the ORF concentration:

$$K_B = \frac{[\text{RNAP}_F] [G_l]}{[\text{RNAP}_l]}, \quad (1.55)$$

¹³This section has been added in this thesis, to account for developments in the formulation post-publication.

which gives the following bilinear constraint:

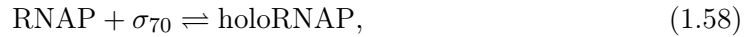
$$K_B \cdot [\text{RNAP}_l] = [\text{RNAP}_F] * [G_l]. \quad (1.56)$$

However, we have seen that G_l is equal to the gene copy number n_l times the DNA concentration, which itself has been discretized into $\sum_{u \in \mathcal{U}} \lambda_u \cdot \text{Dm}_u$. We can thus approximate Eq. 1.56 in the following way:

$$K_B \cdot [\text{RNAP}_l] = n_l \cdot [\text{RNAP}_F] * \sum_{u \in \mathcal{U}} \lambda_u \cdot \text{Dm}_u. \quad (1.57)$$

This expression is a sum of linear variables and products of continuous variables by binary variables, and as such can be linearized using the Petersen linearization scheme. This method can also be trivially extended to ribosomes and their equilibrium with ribosomal binding sites.

Thermodynamic equilibrium of RNAP with promoters and transcription factors Finally, a combination of the strategies detailed in the previous paragraph and post-translational modifications can be used to model the interplay between RNAP and a transcription factor such as the sigma factor 70, σ_{70} . We can model the activation of the RNAP into its holoenzyme (holoRNAP) by σ_{70} with the following reaction:



We can write the binding constant K_B^σ of this reaction using the previous notation system:

$$K_B^\sigma = \frac{[\text{RNAP}] [\sigma_{70}]}{[\text{holoRNAP}]}, \quad (1.59)$$

which gives the following bilinear constraint:

$$K_B^\sigma - \frac{[\text{RNAP}]}{[\text{holoRNAP}]} * [\sigma_{70}] = 0. \quad (1.60)$$

The ratio of holoRNAP to RNAP can be found in publications such as Neidhardt *et al.* (47), and discretized using the methods detailed previously. The product of the discretized ratio and the concentration of σ_{70} can then be linearized using the Petersen transformation. This formulation can be easily adapted for ribosomes and molecules that

bind the translation initiation complex.

Competition for the allocation of expression machinery Several biological systems are impacted by a competitive allocation of expression mechanisms¹⁴. Plasmid insertion in cells, for example, introduces exogenous genetic material that will also use polymerases and ribosomes, or even transcription factors, to express its product. These considerations can also be extended to parasitic interactions, such as viruses using the gene expression machinery of the cell. Since polymerases and ribosomes are limited resources of the cell that are necessary to its growth, their competitive usage will have a negative impact on the cell replication, as shown by Peretti *et al.* (82). Using a combination the methods detailed above, in particular thermodynamic equilibria for polymerases and ribosomes, and polymerase saturation, it is possible to account for such competition. As a result, the total capacity constraints of these molecules can be rewritten as follows:

$$\sum_{l \in \mathcal{L}^{\text{endo}}} X_l + \sum_{l' \in \mathcal{L}^{\text{exo}}} X_{l'} + X_F - X_{\text{tot}} = 0, \quad (1.61)$$

where X_\star denotes the concentration of the macromolecule of interest (polymerase, ribosome, transcription factor) allocated to the expression of the gene \star , $\mathcal{L}^{\text{endo}}$ is set of genes endogenic to the host system, \mathcal{L}^{exo} is the set of exogenic genes from the foreign genetic material, X_F is the concentration of free macromolecules, and X_{tot} the total concentration of the macromolecule. If binding constants are known, the thermodynamic considerations mentioned above will add additional linear relationships between X_{tot} , $\sum_{l \in \mathcal{L}^{\text{endo}}} X_l$, $\sum_{l' \in \mathcal{L}^{\text{exo}}} E_{l'}$, and X_F . To avoid over-constraining the model, it might be necessary to relax Eq. 1.61 in favor of these linear relationships.

1.4.19 Thermodynamics-based constraints

Thermodynamics flux analysis (TFA) (35, 36) imposes constraints on a FBA problem to couple reaction directionality to the standard free energy of reactions and metabolite concentrations. We also introduce constraints that couple the sign of the Gibbs energy of a reaction to its directionality through the use of integer variables and a mixed-integer linear coupling formulation. This framework reduces the feasible flux space and improves the predictive power of FBA by removing thermodynamically invalid flux profiles.

Considering c_i is the concentration of i^{th} metabolite, we define C_i as its scaled logarithm

¹⁴This section has been added in this thesis, to account for developments in the formulation post-publication.

with respect to c_0 so that in standard conditions $c_0 = 1 \text{ M}$:

$$\forall i, \quad C_i = \ln \left(\frac{c_i}{c_0} \right). \quad (1.62)$$

We use the group contribution method (87) to directly calculate $\Delta_r G_j'^o$, the Gibbs energy in solution of the j^{th} reaction. The calculated energy is the net change in the energies of formation of the compounds, which is simply the algebraic sum of the bonds that are broken and formed. This allows to minimize the estimation error of $\Delta_r G_j'^o$, as there is no error coming from the groups that do not react. Hence, we obtain the additional variables:

$$C_i^{\min} \leq C_i \leq C_i^{\max}, \quad (1.63)$$

$$\Delta_r G_{j,\min}'^o \leq \Delta_r G_i'^o \leq \Delta_r G_{j,\max}'^o, \quad (1.64)$$

$$\Delta_r G_{j,\min}' \leq \Delta_r G_i' \leq \Delta_r G_{j,\max}'. \quad (1.65)$$

Some metabolites are not fully characterized, *e.g.* metabolites with -R groups such as fatty acids, or metabolites attached to a Coenzyme A or acyl-carrier protein. In these cases, the group contribution method allows to directly calculate the net change in the standard Gibbs energy. Since these -R groups are often conserved in the reaction, their contribution terms cancel out when calculating the Gibbs energy of the reaction.

The concentration variables are bounded by experimental measurements or physiological assumptions, and the standard Gibbs energies are bounded by the measurement or estimation error. Since the net flux of each reaction has already been split between forward flux (v_j^+) and backward flux (v_j^-), (see Eq. 1.13), we can directly add the constraints described in (35):

$$\Delta_r G_j' - \text{RT} \sum_{i=1}^m \eta_i^j C_i - \Delta_r G_i'^o = 0, \quad (1.66)$$

$$\Delta_r G_j' - K + K \cdot b_j^+ \leq 0, \quad (1.67)$$

$$-\Delta_r G_j' - K + K \cdot b_j^- \leq 0, \quad (1.68)$$

$$v_j^+ - K \cdot b_j^+ \leq 0, \quad (1.69)$$

$$v_j^- - K \cdot b_j^- \leq 0, \quad (1.70)$$

$$b_j^+ + b_j^- \leq 1. \quad (1.71)$$

R denotes the ideal gas constant, T is the temperature in Kelvin, and η_i^j represents the stoichiometry of the metabolite i in the reaction j . K is a big-M constant (bigger than all upper bounds), and b_j^\pm are binary variables. Eq. 1.66 defines the actual Gibbs energy of the reaction as a function of its standard Gibbs energy and the scaled logarithms of metabolite concentrations. Eq. 1.67 and Eq. 1.68 ensure that $\Delta_r G'_j \leq 0 \iff b_i^+ = 1$ and $\Delta_r G'_j \geq 0 \iff b_i^- = 1$. These binary variables are used to block flux in Eq. 1.69 and 1.70 if the thermodynamics do not favor it. Finally, Eq. 1.71 is added to enforce that only one direction is chosen.

1.4.20 Data

mRNA degradation rates constants k_{deg} were taken from Bernstein *et al.* (88). We converted the reported half lives into rate constants using the classical relationship $k = \frac{\log(2)}{t_{1/2}}$. Proteins were approximated to have a half life of 20 h (BNID 111930, (89)).

Catalytic rate constants k_{cat}^j were obtained from Davidi *et al.* (90) for homomeric enzymes. Complex formation reactions for non-homomer enzymes were taken from the supplementary information of O'Brien *et al.* (40) and Lloyd *et al.* (41). EC numbers were obtained from BiGG (91) and the iJO1366 publication (48). Their corresponding k_{cat} values were assigned using conservative (max) values from SabioRK.

Homomer compositions were obtained from Davidi *et al.* (90). Other peptide compositions of enzymes were taken from the supplementary information of O'Brien *et al.* (40) and Lloyd *et al.* (41). Additional information was obtained from the Metacyc/Biocyc database (65, 92) using specialized SmartTables queries (93).

1.4.21 Model modification

The initial model was subjected to minor changes to accommodate for ETFL modeling. In particular, we added:

- Selenocysteine as a metabolite.
- Cysteine to selenocysteine conversion as a pseudo reaction.
- Replacements for the tRNA metabolites and their charging reaction, as dilution has to be considered.

We also modified the biomass reaction by removing its nucleotide and amino acid components, since they are already taken into account by the expression problem as explained in the section Biomass reaction synthesis and mass balance.

1.4.22 Enzyme estimation

Given a reaction in the model, if no enzyme is supplied but the reaction possesses a gene reaction rule, it is possible to infer an enzyme from it. The rule expression is expanded, and each term separated by an **OR** boolean operator is interpreted as an isozyme, while terms separated by an **AND** boolean operator are interpreted as unit peptide stoichiometric requirements. The enzyme is then assigned an average catalytic rate constant and degradation rate constant.

1.4.23 Essentiality analysis

The method for testing gene essentiality in FBA is to evaluate for each reaction the gene-protein-reaction association rules (GPRs) containing the gene of interest. The GPR is a boolean expression where the symbols represent whether a gene is expressed. **OR** operators represent isozymes, and **AND** operators the assembly of several peptides in a complex. To knock a gene out, its symbol in each GPR is simply assigned the value **False**. The GPR of all reactions is subsequently evaluated, and the reactions whose GPR evaluates to **False** are set to have a net flux of 0. Knocking a gene out in ETFL works differently: we replace GPRs with mass balances, and the direct interaction between gene transcription, peptide translation, enzyme assembly, and metabolism. In this context, knocking-out a gene is done by forcing its transcription rate to 0. Indeed, gene-reactions relationships are conveyed directly through the direct contribution of the relevant peptides either as components of the enzyme complex (**AND** operator in GPRs) or as isozymes (**OR** operator). An advantage of this formulation is that it can be used in strain design strategies to optimize directly for knock-outs in a single optimization problem.

If a knocked-out gene does not have enzyme associated with it (because of the lack of composition or k_{cat} information), there will be no catalytic constraint associated with the corresponding enzyme. The absence of catalytic constraint will prevent the reaction to be knocked-out. Hence, because of the missing information, gene essentiality information will be lost. An example is the essential reaction Sulfite reductase NADPH2 (SULR). iJO1366 provides a GPR describing a complex needing b2763 and b2764. The ETFL source (the cobraME model and YeastCyc) could not provide the stoichiometry of the peptides to form the complex, and thus no enzyme is associated to this reaction in the vETFL model. iJO1366 correctly predicts the genes b2763 and 2764 as essential, but ETFL fails because these genes are not associated to any enzyme. As more enzyme data is added to the model, the false positive rate decreases, as we show in the section Essentiality analysis.

For increased performance, the essentiality analysis was cast into a feasibility problem. We put a lower bound on growth equal to 10% of the predicted ETFL growth and set the objective to 0. With this method, essential genes will cause the problem to be infeasible, while non-essential genes will return a feasible solution satisfying at least 10% of the

growth. This method achieved up to a 5-fold reduction in solving time on the most complex models.

1.4.24 Hardware

Computations were done on a 64-bit Ubuntu 18.04.1 LTS (Bionic Beaver); $2 \times$ Intel(R) Xeon(R) CPU E5-2667 v3 @ 3.20 GHz (8 cores, 16 threads per socket); 4×16 Go @ 2400 MHz RAM. Code was run on Python 3.6 on Docker (18.09.0) containers based on the official python 3.6-stretch container, available on ETFL GitHub and ETFL GitLab.

1.5 Code availability

The code has been implemented as a plug-in to pyTFA (94), a Python implementation of the TFA method. It uses COBRApy (95) and Optlang (96) as a backend to ensure compatibility with several open-source (GLPK, scipy, ...) as well as commercial (CPLEX, Gurobi, ...) solvers. We rely on the Python package Biopython (97) for transcribing and translating sequences of nucleotides and amino acids.

The code used to generate the models is freely available under the APACHE 2.0 license at <https://github.com/EPFL-LCSB/etfl> and <https://gitlab.com/EPFL-LCSB/etfl>.

1.6 Data availability

All the data used to conduct this study is available in the `organism_data` subfolder of the repositories. Some of the data has been obtained from publications, for which all the references are provided in the main text, and a copy has been included in our repositories that mentioned above. The code also contains comments crediting the publications from which datasets and values have been obtained.

S1 From biochemistry to constraints. Derivation and formulation. Step-by-step formulation of the biochemistry, from catalytic constraints to transcription.

Supplementary Note 1 Nondimensionalization. Derivation and formulation. Details on the variables and constraint transformations to scale the model.

Supplementary Table 1 Example EP constraint matrix. Representation of the constraint matrix of the EP for a vETFL of iJO1366. Colored cells represent non-zero blocks. Uncolored cells are zero blocks.

Supplementary Note 2 SOP for creating an ETFL model. Tips and Prerequisites. List of required and optional inputs to transform a genome-scale model into an ETFL model

Supplementary Note 3 Note on steady-state assumptions. Justification and details. Detailed account of the assumptions made with respect to steady state and dilution for ME-models

Supplementary Note 4 ETFL Optimization problem. Optimization problem definition. ETFL bilinear formulation and integer-linearized formulation.

Supplementary Data Gene essentiality. vETFL vs iJO1366. List of mismatches in the essentiality of genes between iJO1366 predictions and vETFL predictions.

Supplementary Note 5 Glossary Definitions. Details on technical terms relevant to this interdisciplinary work.

Acknowledgements


The authors would like to thank Prof. Jens Nielsen and Dr. Ibrahim El-Semman for the valuable discussions about the formulation; and Dr. Kaycie Butler for her valuable input on the wording and structure of this manuscript. This work has received funding from the European Union's Horizon 2020 Research and Innovation Programme under the Marie Skłodowska-Curie grant agreement No 722287, the European Union's Horizon 2020 Research and Innovation Programme under grant agreement No 686070, and the Swiss National Science Fund (SNSF) under the grant agreement No 200021_188623.

Author Contributions

P.S. and V.H. designed the formulation and the studies. P.S. wrote the ETFL code, curated the models, and performed the studies. P.S. and V.H. wrote the manuscript.

Competing Interests

The authors declare no competing interests.



2. Dynamic ME-models: Emergence of diauxie as an optimal growth strategy under resource allocation constraints in cellular metabolism

Chapter 2. Dynamic ME-models: Emergence of diauxie as an optimal growth strategy under resource allocation constraints in cellular metabolism

Pierre Salvy¹, Vassily Hatzimanikatis^{1,*},

¹ Laboratory of Computational Systems Biotechnology, École Polytechnique Fédérale de Lausanne (EPFL), Lausanne, Switzerland

* Corresponding author: vassily.hatzimanikatis@epfl.ch

This second chapter is a nod to the French biologist Jacques Monod. Jacques Monod discovered that, when several sugars are at its disposition, the bacterium *E. coli* consumes them in a specific order — a behavior he called diauxie. Current computational models need specific assumptions to be able to accurately reproduce this behavior. Using a novel state-of-the-art modeling framework, I show diauxie can be explained simply as an optimal behavior under constraints on the protein amount in a cell. The method allows a dynamic description of the physiology of diauxie, at the proteome level. I validate the model by reproducing experimental results, and successfully predict a diauxic behavior on a growth medium containing two types of sugar, glucose and lactose. Finally, I claim that the regulation mechanism inducing diauxie (the *lac* operon) is a control system to implement growth optimality at the cellular level.

The chapter is adapted from the preprint P. Salvy and V. Hatzimanikatis, “Emergence of diauxie as an optimal growth strategy under resource allocation constraints in cellular metabolism,” *bioRxiv*, 2020, which has been submitted as an article to a peer-reviewed journal. Vassily Hatzimanikatis and I worked on the formulation, and designed the studies to perform. I performed the literature search and data curation, with input and guidance from Vassily Hatzimanikatis. I wrote all the code to implement the formulation, and the scripts to perform the studies. I curated the data to make the models, and made the figures.

All the code and documentation is available under the APACHE 2 license, in the subdirectory `work/detfl` at:

<https://github.com/EPFL-LCSB/etfl>

<https://gitlab.com/EPFL-LCSB/etfl>

The content of this chapter is also available as a preprint on bioRxiv at

<https://doi.org/10.1101/2020.07.15.204420>.

2.1 Introduction

In his pioneering work on the growth of bacterial cultures, the French biologist Jacques Monod (99) observed that the growth of *Escherichia coli* (*E. coli*) in a mixture of carbohydrates followed two distinct exponential curves separated by a plateau — a phenomenon he called diauxie. Hypothesized to allow optimal growth of the culture (100), this cellular behavior corresponds to the sequential consumption of sugars, where one sugar is preferentially consumed, and the second is consumed after depletion of the first. Although current optimality-based computational models can predict diauxie, these lack a detailed description of protein dynamics during the phenomenon (101). Diauxie is an evolved, complex behavior, and its occurrence is controlled by the regulation network of the *lac* operon in *E. coli* (102, 103). The emergence of such a control mechanism is the product of evolutionary pressure, and being able to fully elucidate its *raison d'être* in terms of cell physiology is an important milestone to understand and better engineer the intracellular dynamics of bacterial growth. There is thus a need for a formulation describing diauxie at the proteome level.

Genome-scale models of metabolism (GEMs) combine constraint-based modeling and optimization techniques to study cell cultures (27, 104, 105). A key method for studying GEMs is flux balance analysis (FBA) (28), which formulates a linear optimization problem that employs stoichiometric constraints through the mass conservation of metabolites given their synthesis and degradation reactions. Under the typical steady-state and growth-rate maximization assumptions, FBA models predict the simultaneous consumption of two or more carbon sources to achieve the maximum possible growth (101). However, this contradicts Monod's observation of distinct, sequential phases of carbon consumption and suggests that diauxie does not come from stoichiometric constraints.

To account for diauxie beyond stoichiometric modeling, we looked into other biological features. In his review on catabolite repression, Ullmann (100) reports a remark from Magasanik (106) on the limited size of the enzyme pools in the cell. Magasanik argues that catabolite repression prevents the synthesis of specific enzymes, thus preventing a surcharge of the cell proteome. Indeed, a cell has a physiological constraint on the amount of enzymes it can contain, or proteome allocation constraint. It is reasonable to expect that under such allocation constraints, the system will preferentially distribute its now-limited catalytic capacity towards pathways utilizing the most efficient substrate/enzyme combination. In this respect, it appears models that account for proteome limitation in cells may be able to account for diauxie. Towards this end, the role of protein limitation in diauxie was demonstrated by Beg *et al.* (37) with their formulation of FBA with molecular crowding (FBAwMC). Their method correctly predicts the uptake order of five different carbon sources in a batch reactor, using a proteome allocation constraint. In a push towards more global models, models of metabolism and expression (ME-models) (39, 40) include proteome allocation, but also gene expression mechanisms, a modeling paradigm that is ideal for studying diauxie at the proteome level. ME-models also fully

Chapter 2. Dynamic ME-models: Emergence of diauxie as an optimal growth strategy under resource allocation constraints in cellular metabolism

describe the requirements of enzyme synthesis, degradation, and dilution effects, as well as mRNA and enzyme concentrations.

Since the diauxie is also a time-dependent phenomenon, we chose to complement ME-models with a dynamic modeling approach. Dynamic FBA (dFBA) (64) is a generalization of FBA for modeling cell cultures in time-dependent environments. In its original static optimization approach (SOA) formulation, the time is discretized into time steps, and an FBA problem is solved at each step. At each iteration, kinetic laws and the FBA solution are used to update the boundary fluxes, extracellular concentrations, and cell concentration, based on the amount of substrate consumed, byproducts secreted, and biomass produced by the cells. We expected that the combination of a dFBA and ME-models would yield a formulation that can describe diauxie at the proteome level.

However, we identified three major challenges in the conception of dynamic models of metabolism and expression. First, while dFBA studies of metabolic networks can be solved by common linear solvers, ME-models are non-linear by nature, and significantly more complex. The new species and reactions introduced and considerations of the interactions between enzyme expression and metabolism result in nonlinear problems that are often 1–2 orders of magnitude bigger in terms of constraints and variables than the corresponding linear (d)FBA problem. The increase in complexity is compounded when iteratively solving an optimization problem. As a result, combining ME-models and dynamic studies brings along difficulties that arise from the high computational cost of solving multiple times, with different conditions, these large, non-linear problems. Second, the use of iterative methods presents the additional challenge of alternative solutions, which can span several physiologies. It is thus necessary to find, for each time step, a suitable representative solution that will be used to integrate the system. This also poses the problem of finding a set of initial conditions for the system. Third, the current state-of-the-art models present limitations at the proteome level. Lloyd *et al.* (41) developed an efficient ME-model for *E. coli*, and Yang *et al.* used it to formulate a dynamic analysis framework (dynamicME) (107) similar to dFBA. However, the assumptions introduced to alleviate the computational complexity of their model limit some aspects of the modeling capabilities of their method (Supplementary Note S1). In particular, DynamicME forces a strict coupling between enzyme concentrations and fluxes. However, a change in the growth conditions will trigger a change in the proteome allocation to adapt to a new metabolic state, or lag phase. During that time, it is expected that some previously active enzymes will not be able to carry flux in the new conditions. Therefore, enzyme flux and concentration will decouple, unless the enzyme composition of the proteome changes at the same rate as the environment. As a result, the method cannot simulate lag phase during glucose depletion and proteome reallocation.

Both dynamic models and models including gene expression mechanisms are important components in the development of successful predictive biology (42). We propose a dynamic method which tackles the challenges mentioned above and models diauxie at

the proteome level. To this effect, we used our recently published framework for ME-models, ETFL (43). The formulation of ETFL permits the inclusion of thermodynamics constraints in expression models, as well as the ability to describe the growth-dependent allocation of resources. ETFL is faster than previous ME-model formulations, thanks to the use of standard mixed-integer linear programming (MILP) solvers (43). We herein leverage ETFL for dynamic analysis, in a method called dETFL. It includes a method based on Chebyshev centering to robustly select a representative solution from the feasible space at each time step. The representative solution captures phenotypic and genotypic differences between cells precultured in different media. (d)ETFL solves the problem of computational accuracy lacking in previous models by performing a systematic scaling of its constraints, eliminating the need for dedicated solvers. This allows models to be solved efficiently, without resorting to a strict coupling of enzymes and fluxes. As a result, whole-proteome reconfiguration during sugar consumption can be simulated, which will enable the modeling of the lag phase in diauxie.

Herein we model the emergence and dynamics of diauxie arising at the proteome level. We first propose a small conceptual model of a cell, with a limited proteome, and demonstrate its ability to predict diauxie under a minimal set of assumptions. Using the dETFL method, we subsequently show these assumptions hold in *E. coli*, and reproduce experimental results of bacterial growth. Finally, we apply the dETFL framework to the growth of *E. coli* in a glucose/lactose mixture in a batch reactor, and demonstrate that it robustly predicts diauxie as well as the preferential consumption of glucose over lactose. Overall, dETFL offers a method to robustly survey intracellular dynamics of cellular physiology under changing environmental conditions.

2.2 Results

2.2.1 Conceptual model for the emergence of the diauxie phenotype from proteome limitation

We designed a simplified conceptual model, to illustrate diauxie from proteome limitations, as described in Fig.2.1-a. The model includes both glucose and lactose as substrates, and it is a simplified version of the *E. coli* metabolism based on four considerations:

- (C1) The biomass carbon yield on glucose is slightly higher than that of lactose (108)
- (C2) Glucose and lactose are taken up and converge to a common intermediate metabolite, glucose 6-phosphate (G6P). Glucose is transformed into G6P by a glucokinase. The lactose pathway (Leloir pathway) splits the lactose, a disaccharide, into its glucose and galactose subunits. The galactose is then converted to G6P by a series of enzymes.
- (C3) The Leloir pathway requires one enzyme to split the lactose into glucose and

Chapter 2. Dynamic ME-models: Emergence of diauxie as an optimal growth strategy under resource allocation constraints in cellular metabolism

galactose, four enzymes to convert galactose into glucose-1-phosphate (109, 110), and one to convert glucose-1-phosphate into G6P; this brings the total to six enzymes needed for the synthesis of two G6P, which is equivalent to three enzymes per G6P;

- (C4) The molecular weight of each of the enzymes in the lactose pathway is around 60-90 kDa (111), which is heavier than the 33 kDa glucokinase (Uniprot ID A7ZPJ8).

Based on these considerations, we devised a conceptual model of glucose and lactose metabolism for *E. coli*. The model accounts for the consumption of the two substrates, which both synthesize an intermediate metabolite that is then used to make biomass. We thus made five modeling assumptions:

- (A1) Glucose has a slightly higher carbon yield than lactose — based on (C1).
(A2) The glucose and lactose metabolism leading to the intermediate G6P are catalyzed by two different enzymes — based on (C2).
(A3) The molecular weights of the enzymes are the same, and three times more enzymes are required for lactose metabolism than for glucose metabolism — based on (C3).
(A4) The catalytic activities of the two enzymes synthesizing G6P are similar.
(A5) The variation of enzyme concentrations reaches a maximum at each time step.
(A6) The total enzyme amount in the cell is limited.

The mathematical formulation of the problem (Fig.2.1-b) involves one mass balance, one conservation equation of the total enzymes, two inequalities that constrain the metabolism for glucose and lactose as a function of the corresponding enzymes concentrations, and two enzyme variation constraints. Due to total enzyme conservation, the two maximum activity constraints are not independent. This constraint is similar to that found in other approaches accounting for proteome allocation such as FBAwMC (37).

The conceptual model is able to predict diauxic behavior in our system. The model shows the preferential consumption of glucose over lactose (Fig.2.1-d), controlled by a switch in the proteome composition over time (Fig.2.1-c). The diauxic phenomenon is due to the fact that the system will invest all the (limited) enzyme resources into the metabolism of glucose, which is both the highest yielding substrate ((C1), (A1)), and the one with least enzyme requirements ((C3) and (A3)). As the glucose is depleted, the uptake flux is reduced and the system gradually allocates part of its proteome for enzymes needed for lactose metabolism. This gradual proteome reallocation corresponds to the observed lag phase in an experimental system. While this conceptual model lacks catabolite repression

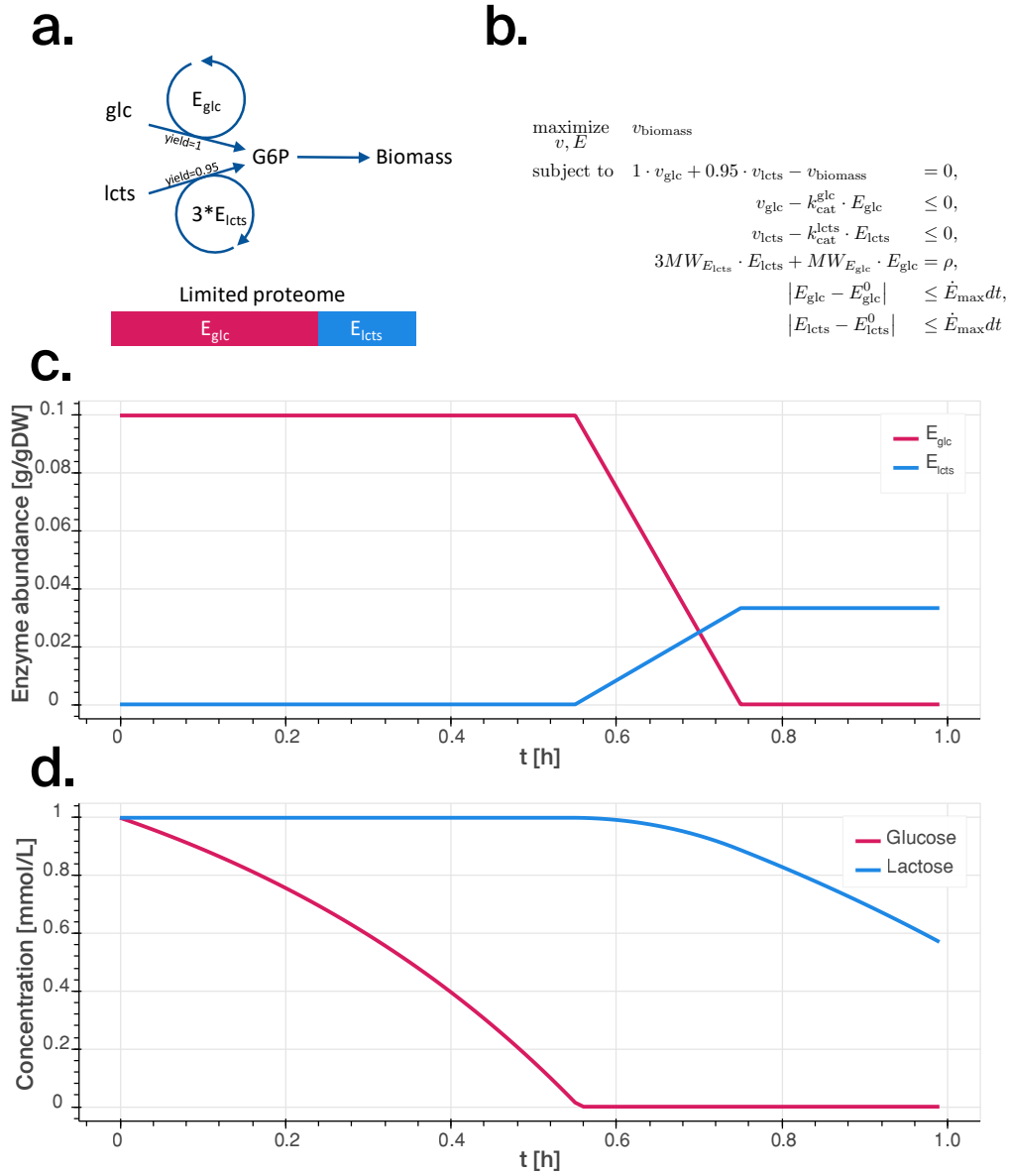


Fig. 2.1. **a.** Conceptual model used for the preliminary analysis, where “glc” stands for glucose, “lcts” for lactose. The catalytic efficiency of the enzymes are assumed to be the same. Three enzymes are assumed to be necessary to produce the intermediate metabolite G6P from lactose, and only one enzyme is required from glucose. **b.** Optimization problem used to represent the model. v are fluxes, E are enzyme concentrations, E^0 are reference values, MW are molecular weights, ρ is the mass fraction of the cell occupied by the enzymes we consider, \dot{E}_{max} is the maximal variation of enzyme concentration over time, and dt is the integration interval. **c.** Enzyme content over time for the conceptual model growing on a mixed substrates. **d.** Changes in sugar content of the batch reactor over time.

Chapter 2. Dynamic ME-models: Emergence of diauxie as an optimal growth strategy under resource allocation constraints in cellular metabolism

mechanisms, it can still describe diauxie phenomenon from only the proteome capacity constraint.

The lack of proteome limitation and enzyme catalytic constraints is why the original FBA approach fails to predict diauxie, which leads to the simultaneous utilization of both sugar substrates. However, another important constraint to accurately describe the lag phase is the limitations on the rate of change in enzyme concentrations, without which the proteome switch would occur instantaneously (see Supplementary Figure S9). A cell needs time to adapt its proteome, and the limits on the rate of change of enzyme concentrations represent the catalytic limitation of the cell to break down old enzymes and synthesize new ones better adapted to the new conditions.

The conceptual model also allows for the study of the conditions under which the system switches to lactose as a carbon source, and the identification of parameters responsible for this behavior. If we note respectively for glucose and lactose, the specific growth rates on each substrate $v_{biomass}^{glc}$, $v_{biomass}^{lcts}$, the carbon yields Y_{glc} , Y_{lcts} , and the catalytic rate constants k_{cat}^{glc} , k_{cat}^{lcts} of enzymes at concentrations E_{glc} , E_{lcts} , then the preferred carbon source will change to lactose if and only if:

$$v_{biomass}^{glc} < v_{biomass}^{lcts}, \quad (2.1)$$

$$Y_{glc} \cdot v_{glc}^{max} < Y_{lcts} \cdot v_{lcts}^{max}, \quad (2.2)$$

$$Y_{glc} \cdot k_{cat}^{glc} \cdot E_{glc}^{max} < Y_{lcts} \cdot k_{cat}^{lcts} \cdot E_{lcts}^{max}. \quad (2.3)$$

If the amount of available enzymes is represented by ρ , as a fraction of the total cell mass (in g.gDW⁻¹), and assuming different molecular weights MW_E , the proteome limitation constraint will be written:

$$MW_{E_{glc}} \cdot E_{glc} + 3MW_{E_{lcts}} \cdot E_{lcts} = \rho. \quad (2.4)$$

The maximal achievable values for the enzyme concentrations will be $E_{glc}^{max} = \rho / MW_{E_{glc}}$ and $E_{lcts}^{max} = \rho / (3MW_{E_{lcts}})$. Replacing these values in Eq. 2.3 directly gives the condition :

$$\frac{Y_{glc}}{Y_{lcts}} < \frac{k_{cat}^{lcts}}{k_{cat}^{glc}} \cdot \frac{MW_{E_{glc}}}{3 \cdot MW_{E_{lcts}}}. \quad (2.5)$$

In our conceptual model, $MW_{E_{glc}} = MW_{E_{lcts}}$, and we can simplify Eq. 2.5:

$$3 \cdot \frac{Y_{glc}}{Y_{lcts}} < \frac{k_{cat}^{lcts}}{k_{cat}^{glc}}. \quad (2.6)$$

This expression identifies the boundary in the parameter space that separates the preferential use of glucose versus lactose.

These calculations can be generalized for a more realistic model, by accounting for the molecular weight of the enzymes and setting an adequate proteome fraction allocated to carbon metabolism. In practice, the catalytic efficiencies of the glycolytic enzymes are also higher than those of the Leloir pathway ((A3), see Supplementary Table S2), and the Leloir pathway enzymes are heavier ((C4), (A4)) which favors glucose consumption even more. Additionally, we did not consider the synthesis cost of the enzymes used to carry the fluxes in each pathway. Taking such property into account would also strengthen the preference towards glucose, as fewer enzymes are needed for its metabolism.

2.2.2 Diauxie in genome-scale, ME-models with thermodynamic constraints

Going beyond a conceptual model, we next used dETFL to model diauxie in a ME-model of *E. coli*. This method allowed us to study metabolic switches in response to a changing environment, under the aspect for intracellular enzyme and mRNA concentrations. To do this, we studied how ME-models can describe diauxie in experiments where *E. coli* are grown in two different conditions. Firstly, we investigated the growth of *E. coli* on glucose. In this experiment, the cell exhibit overflow metabolism, or the secretion of acetate, even under aerobic conditions. Experimentally, the bacterium reutilizes the secreted acetate after glucose depletion, a form of diauxic behavior. This type of study was also used as the first proof of concept for dynamic FBA (64) Thus, we first validated the dETFL model by demonstrating its ability to model a first diauxic phenotype: overflow metabolism and acetate secretion in the presence of excess glucose, followed by acetate reutilization on glucose depletion. Secondly, we reproduced Jacques Monod’s experiment of the diauxic growth of *E. coli* in an oxygenated batch reactor (99) with a limited carbon supply made of a mixture of glucose and lactose. We aimed at reproducing the results shown in the conceptual model on a model of a real organism, and characterize the intracellular dynamics underlying the glucose/lactose diauxic behavior.

To conduct these studies, we used the *E. coli* model published by Salvy *et al.* that is based on the genome scale model by Orth *et al.* iJO1366 (48), and was assembled using ETFL. This model is significantly bigger than the conceptual model studied in the previous section, with 5295 species, 8061 reactions and 578 enzymes. A summary of the model is available in Table 2.1.

For the integration of the dynamic method, it is important to choose a time step that respects the quasi-steady-state assumptions on which the FBA and ETFL frameworks depend (43). We used a time step of $0.05 \text{ h} = 3 \text{ min}$ for the numerical integration, as this is around ten times smaller than a typical doubling time for *E. coli*, and efficiently

Chapter 2. Dynamic ME-models: Emergence of diauxie as an optimal growth strategy under resource allocation constraints in cellular metabolism

Table 2.1. Properties of the vETFL model generated from iJO1366.

Growth upper bound $\bar{\mu}$	$3.5h^{-1}$
Number of bins N	128
Resolution $\frac{\bar{\mu}}{N}$	$0.027h^{-1}$
Number of constraints	69323
Number of variables	50010
Number of species	5295
– Metabolites	1806
– Enzymes	578
– Peptides	1433
– mRNAs	1433
– tRNAs	21×2
– rRNAs	3
Number of reactions	8061
– Metabolic	1840
– Transport	733
– Exchange flux	330
– Transcription	1433
– Translation	1433
– Complexation	578
– Degradation	2011
Number of metabolites $\Delta_f G'^o$	1737
Number of reactions $\Delta_r G'^o$	1787
Percent of metabolites $\Delta_f G'^o$	93.9%
Percent of reactions $\Delta_r G'^o$	69.5%

balances the integration approximation and solving time.

Diauxic growth on glucose and acetate We compared the accuracy of our computational modeling of diauxie to experimental findings. Specifically, we studied the diauxic growth of *E. coli* on glucose using in batch reactors using experimental data published in Varma *et al.* (27) and Enjalbert *et al.* (112). Previously, Varma *et al.* (27) used their data to validate a stoichiometric model of *E. coli* in quasi-steady state, whereas the data from Enjalbert *et al.* (112) was used to validate a population-based approach of dFBA by Succuro *et al.* (113).

To reproduce the results of these two batch growth experiments, we applied constraints to the uptake of glucose and oxygen in the dETFL model (see Materials and Methods). The initial uptake rate of glucose is set to $15 \text{ mmol.gDW}^{-1}.\text{h}^{-1}$. This value is characteristic of a typical physiology for *E. coli* growing on glucose with excess oxygen (27, 64, 48, 113).

We also matched the initial concentrations of cells, glucose, and acetate are set to the values of the experimental data. Oxygen transfer was considered free (no kinetic law on uptake) in a first approximation, as done by Succuro *et al.* (113).

Our simulations agreed with the published experimental data. The temporal evolution of the glucose and acetate concentrations in the simulated batch reactor agreed with both the Varma and Enjalbert datasets, as shown in Fig. 2.2-a and 2.2-b, respectively. The cell concentration and specific growth rate also follow a similar trend (Fig. 2.2-c and 2.2-d). Both of the simulations predict a first phase where the bacteria grow steadily on glucose, which is sustained until glucose is depleted in the medium. During that time, acetate is steadily secreted by the cell, due the overflow metabolism. When extracellular glucose is depleted, the residual acetate is consumed by the cell. We observe a sharp drop in the cell growth rate, and the simulation ends when no acetate is left in the medium.

We achieved these simulated curves with no fitting. The results are the predictions of dETFL — given only the starting point of the simulation, and then aligning the curves based on the time of glucose depletion to account for experimental lag phases. The discrepancy between the simulation and the experiment data points can be attributed to several factors. First, several simulation parameters, including the maximal uptake rate for glucose, oxygen, and acetate, and the acetate maximal secretion rates, are reported with a 50% variability between the Varma and Succuro studies. We chose a common set of parameters that showed good qualitative agreement with both datasets. Changing these parameters can alter the quantitative behavior of the model, but the models always shows the same two phases. Second, variability in the experimental setup, including the *E. coli* strain, can also account for the difference in the reported glucose uptake rate by their respective authors. ME-models, and ETFL in particular, can account for the strain variability if the genetic differences (gene knock-outs, enzyme activities, enzyme over-

Chapter 2. Dynamic ME-models: Emergence of diauxie as an optimal growth strategy under resource allocation constraints in cellular metabolism

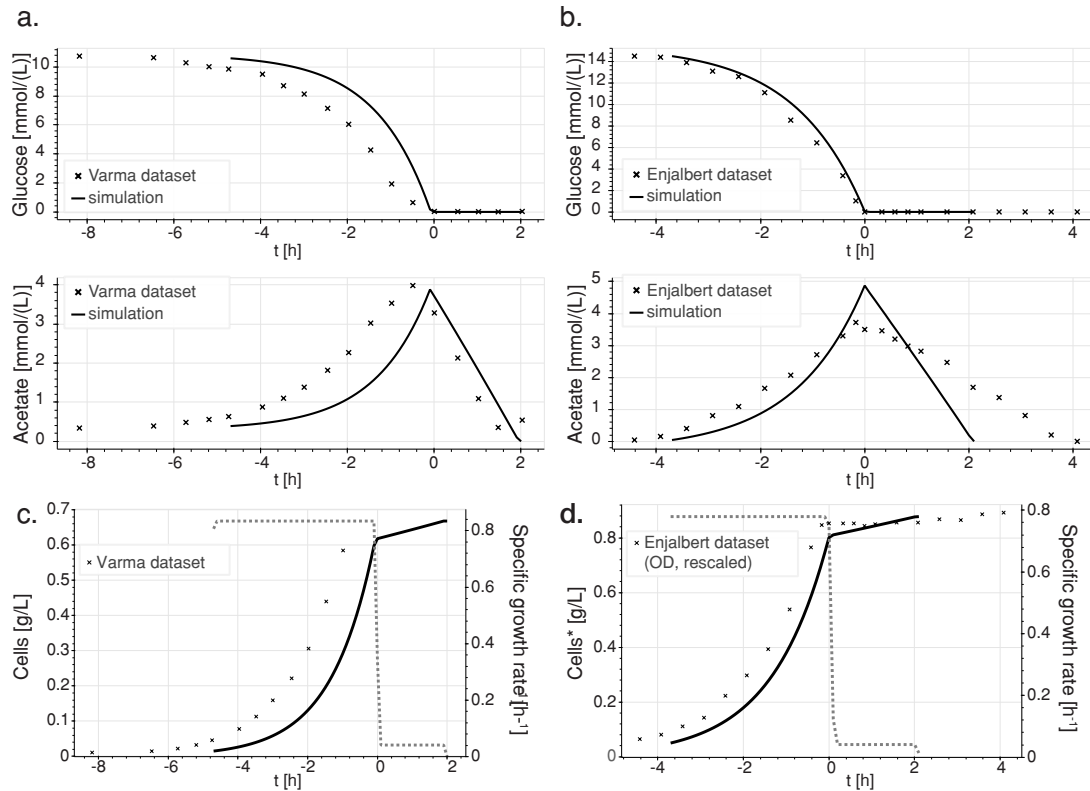


Fig. 2.2. Comparison of simulated and experimental data of glucose depletion over time (Varma and Enjalbert). Simulated data represented by a solid line, experimental data by crosses. **a.** Temporal evolution of the simulated extracellular concentrations of glucose, and acetate (full lines), versus experimental data (Varma dataset) (crosses). **b.** Temporal evolution of the simulated (full lines) extracellular concentrations of glucose, and acetate (solid lines), versus experimental data (Enjalbert dataset) (crosses). **c.** Cell concentration (full line) and growth rate (dashed line) over time, simulation and Varma dataset (crosses). **d.** Cell concentration (solid line) and growth rate (dashed line) over time, simulation and Enjalbert dataset (crosses). * Experimental values were in optical density (OD600), and were linearly scaled to represent cell concentrations.

expression) are known. Overall, these results show the dETFL framework for ME-models is able to reproduce experimental measurements of glucose uptake, acetate secretion, and biomass production in glucose-acetate diauxic growth. Our findings validate the dETFL framework as a modeling method to study the batch growth of single organisms or communities on multiple substrates and suggest its utility for investigating diauxie in mixed-substrate media.

Diauxic growth on glucose and lactose Diauxic experiments show that, on a mixed medium of glucose/lactose, *E. coli* will preferentially consume glucose first, and then lactose (114, 115). Modeling the diauxic growth of *E. coli* with dETFL should capture the lag phases and proteomic reconfiguration that are caused by the shift to a new carbon source. Therefore, this is an ideal system to challenge the ability of ME-models to describe the dynamic reorganization of the bacterial proteome. dFBA will always predict simultaneous uptake of both carbon sources, since it includes no term associated to the proteomic cost of their uptake. In contrast, ME-models describe the synthesis of enzymes, and their contribution to the overall proteome. As a result, ME-models capture the competitive allocation of the proteome to the transport of different carbon sources.

For reference, the pathways related to the glucose and lactose metabolism to G6P are summarized in Fig. 2.3. The figure highlights the multiple additional steps involved in the lactose pathway to form G6P, compared to the shorter glucose pathway.

To initialize the model for the simulation of diauxic growth, we first simulate the pre-culturing in glucose by running the model with the same standard physiology as before, with an uptake rate of $15 \text{ mmol.gDW}^{-1}.\text{h}^{-1}$ for glucose, and no lactose initially present. The model is subsequently run with these initial conditions on a mixture of glucose and lactose, at the physiologically relevant concentrations of 1 mmol.L^{-1} and 2 mmol.L^{-1} , respectively. The cell concentration is set at 0.05 g.L^{-1} . After this initialization step, we ran the simulation according to the method detailed before.

The time evolution of the extracellular metabolite concentrations, cellular exchange fluxes, specific growth rate, and total biomass of the culture exhibit four phases (Fig. 2.4). We observe a first phase similar to the previous experiment, where glucose is taken up at a rapid rate, until its depletion, with the simultaneous production of acetate through overflow metabolism (Fig. 2.4-a and -b). During this phase, the growth rate is steady and high (Fig. 2.4-c). Relative to glucose, lactose is taken up at lower rates (Fig. 2.4-a). In the second phase, the specific growth rate decreases sharply while the proteome reallocates its enzymes for lactose metabolism. We also observe a drop in acetate secretion during the proteome switch and short period of acetate re-consumption. This is the lag phase, where acetate is used as a carbon source while the proteome is reconfigured to metabolize lactose. This reconfiguration shows a reduction of the total mass of enzymes that convert glucose into G6P, and an increase in the total mass of enzymes responsible for the conversion of

Chapter 2. Dynamic ME-models: Emergence of diauxie as an optimal growth strategy under resource allocation constraints in cellular metabolism

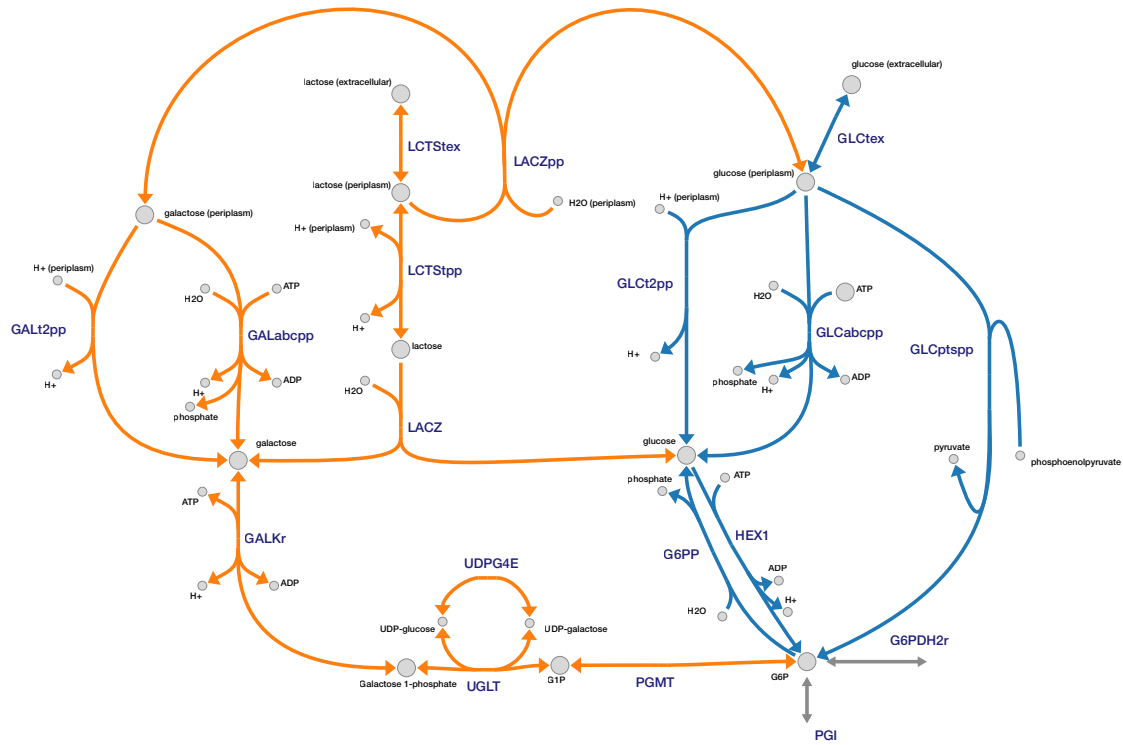


Fig. 2.3. Possible uptake routes for glucose (blue) and lactose (orange), towards glucose-6-phosphate (G6P). The splitting of lactose by LACZ can be done either intracellularly, or in the periplasm. Routes towards the main central carbon metabolism are in gray. Figure made using Escher. (91)

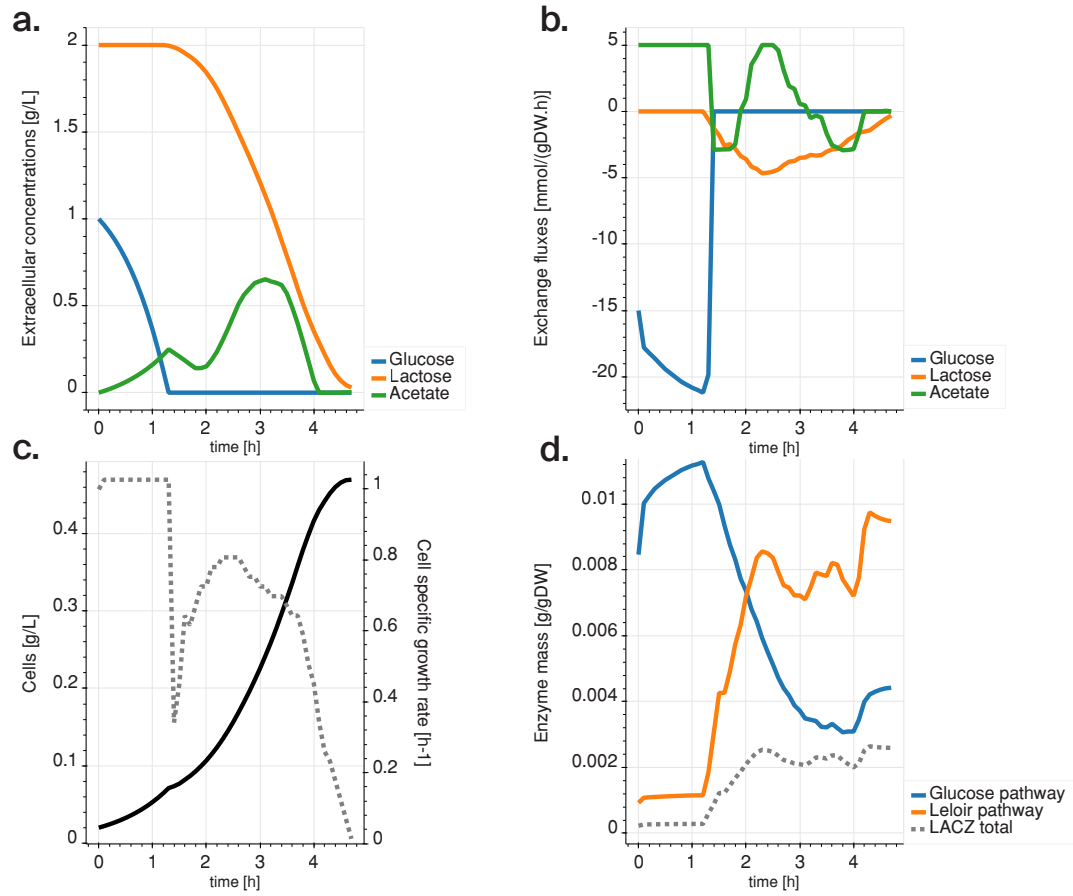


Fig. 2.4. Diauxic simulation with glucose-only preculture: **a.** Temporal evolution of the extracellular concentrations of glucose (blue), lactose (orange), and acetate (green). **b.** Exchange rates of the cell. Positive exchange rates mean production, negative exchange rates mean consumption. **c.** Cell concentration (solid line) and growth rate (dashed line) of the culture over time. **d.** Mass of enzymes allocated to the transformation of glucose (blue) and lactose (orange) in G6P. The dashed gray line shows the levels of β -galactosidase (LACZ) enzyme (in the lactose pathway).

Chapter 2. Dynamic ME-models: Emergence of diauxie as an optimal growth strategy under resource allocation constraints in cellular metabolism

lactose to G6P (Fig. 2.4-d). The third phase is characterized by a peak in lactose uptake and cell growth, followed by a decline as lactose becomes scarce. In the fourth phase, when lactose becomes scarce, the residual acetate is being taken up instead of secreted. Since it happens after lactose uptake falls below a low threshold, it indicates that lactose consumption is preferred to that of acetate. More details on the time-dependent enzyme concentrations of the glucose and Leloir pathways can be found in the Supplementary Figure S3 and S4, respectively.

We next sought to assess the robustness of our diauxie prediction, and to determine whether the delayed utilization of lactose was an artifact of the initial conditions used in the simulation. We conducted a new simulation that included a preculture wherein the *E. coli* model initially only had access to lactose, with an uptake rate of $5 \text{ mmol.gDW}^{-1}.\text{h}^{-1}$. Following this preculture, we ran the simulation with identical initial conditions to previous experiments in terms of glucose, lactose, and cell concentrations.

Over the course of this simulation, we observed the same four phases : (1) preferred glucose consumption; (2) proteome switch with acetate uptake; (3) lactose consumption; and (4) acetate reutilization. Initially, glucose is taken up at a significantly smaller rate than that of the glucose preculture. The glucose uptake rate then gradually increases until glucose depletion (Fig. 2.5-a and -b). Comparatively, the lactose uptake rate stays low while the glucose uptake rate increases during the first phase of the experiment. The evolution of the growth rate is similar to that of the previous experiment (Fig. 2.5-c). Though the model was pre-cultured in lactose, the total amount of enzymes transforming lactose decreases while glucose is available (Fig. 2.5-d). We observe a delay, close to the cell doubling time, for initiating the utilization of glucose compared to the glucose-preculture experiment. We attribute this delay to proteome switch, from a proteome optimized for lactose consumption, to a proteome optimized for glucose consumption in this phase. In the second phase, after glucose depletion, we also observe acetate reutilization, while the enzymes needed for lactose conversion to G6P are resynthesized. In the third phase, the proteome shifts again to accommodate lactose consumption. As a result, the lactose uptake rate increases. In the final phase, acetate reutilization initiates again under scarce conditions. More details on the time-dependent enzyme concentrations of the glucose and Leloir pathways can be found in the Supplementary Figure S5 and S6, respectively.

These simulations show strong qualitative agreement with the experimental data, for both the glucose and the lactose precultured conditions. In particular, Kremling *et al.* (115) showed a similar evolution of extracellular concentrations with a two-phase consumption of sugars. They also demonstrated that intracellular LACZ enzyme levels increase when lactose is the sole substrate left, and decrease when glucose is consumed – even after a lactose pre-culture (Fig. 2.4-d and Fig. 2.5-d). Interestingly, these agreements were achieved without adjusting any of the parameters or settings of the original ME-model. However, a key element for the consistency between the model simulations and the cellular state is a robust accounting of the intracellular states (mRNA species, enzymes and

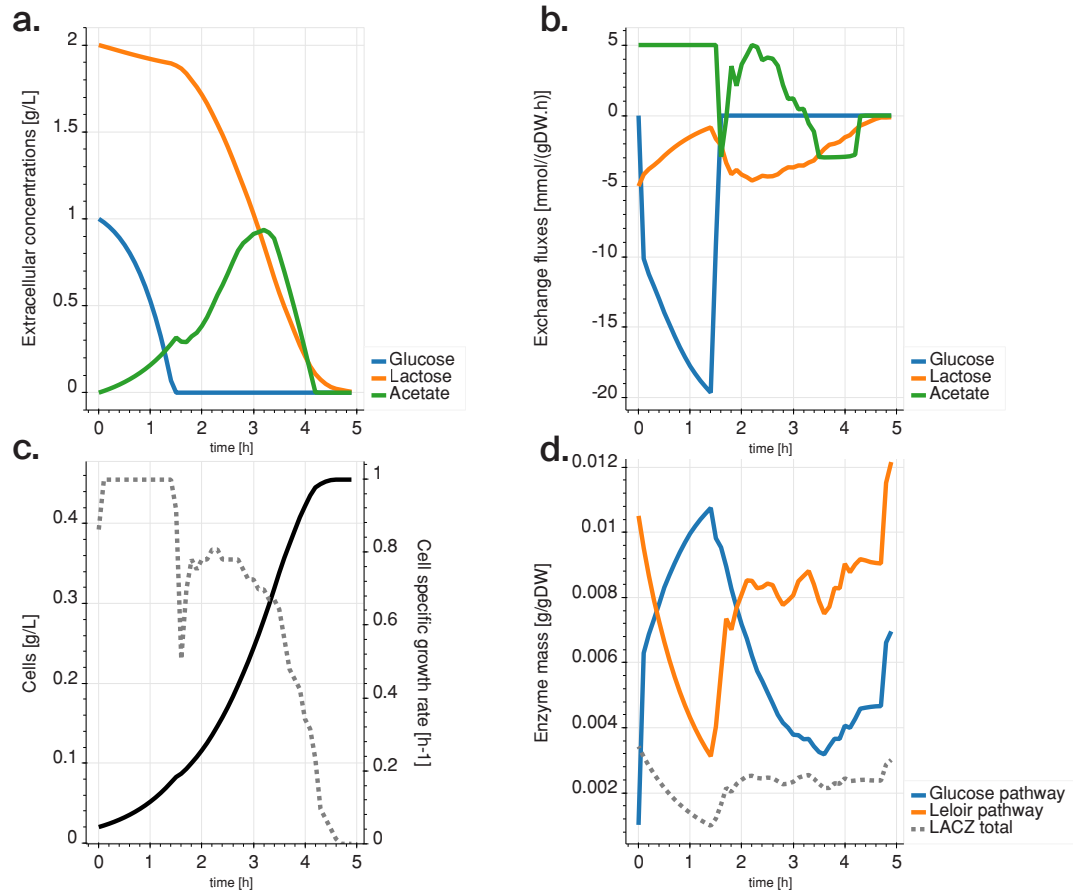


Fig. 2.5. Diauxic simulation with lactose-only preculture: **a.** Temporal evolution of the extracellular concentrations of glucose (blue), lactose (orange), and acetate (green). **b.** Exchange rates of the cell. Positive exchange rates mean production, negative exchange rates mean consumption. **c.** Cell concentration (solid line) and growth rate (dashed line) of the culture over time. **d.** Mass of enzymes allocated to the transformation of glucose (blue) and lactose (orange) in G6P. The dashed gray line shows the levels of β -galactosidase (LACZ) enzyme (in the lactose pathway).

Chapter 2. Dynamic ME-models: Emergence of diauxie as an optimal growth strategy under resource allocation constraints in cellular metabolism

fluxes) between consecutive time steps. This has been made possible by the use of the Chebyshev centering of the cellular states in the dETFL formulation as detailed in the Methods section.

Our results strongly suggest that diauxie in *E. coli* is an optimal growth behavior. Our conceptual study suggests it is the consequence of the maximization of the cell-specific growth rate under the constraint of a limited proteome. This optimal behavior of privileging glucose consumption over lactose does not come from the pre-culturing step, but instead from the optimality of the system itself under the constraint of proteome allocation for sugar consumption. We performed additional studies and demonstrated that this behavior is not due to differences in enzyme catalytic efficiencies between the two pathways, as switching the k_{cat} values does not change the trend (Supplementary Figure S10). Finally, we showed that the lag time observed in experiments is determined by the proteome reallocation, and quantitatively predicted changes in the amount of enzyme for each pathway.

2.3 Discussion

We devised both a conceptual model and a dynamic ME-model which reproduce a diauxic behavior in *E. coli*, a phenomenon that cannot be captured with current state-of-the-art models. From simulation, we determined that the preferential consumption of glucose over lactose in *E. coli* is a combined effect of its limited proteome size, enzyme properties, and substrate yield. Our model demonstrates, at the proteome level, the mechanisms of the proteome switch between conditions, and provides a method to resolve the intracellular dynamics of bacterial growth. In agreement with experimental observations, our model predicts a diauxic behavior on a medium of mixed of sugars.

In our simulations, we observed lag phases concurrent with proteome switching. The co-occurrence of the proteome reallocation and acetate reutilization suggests secreted acetate can work as an energy reserve and help the cell adapt to changing environmental conditions. The dETFL model was also able to capture different dynamic trajectories in cell fates that were dependent on the pre-culture conditions.

The preferential consumption of one carbon source vs the other is the result of an optimal trajectory of the system under the constraints of mass-balance, resource allocation and thermodynamics. These constraints are directly connected to the chemistry of the metabolic pathways in bacteria. Our conceptual model suggests that the diauxic phenomenon might be controlled through the engineering of three aspects: (i) the specific activity of enzymes (k_{cat}), (ii) the molecular weight of the enzymes, and (iii) the number of steps involved in the substrate metabolism. The molecular weight and activity of enzymes can be altered through protein engineering, and alternative chemistries from heterologous pathways provide avenues for modifying substrate metabolism (116).

While dETFL does not account for catabolite repression, it can quantitatively describe the behavior of a cell operating under the influence of the *lac* operon. Our results imply that the genetic circuits responsible for catabolite repression are evolved as a controller to implement robust dynamic control of the optimal growth. In this regard, the catabolite repression through the *lac* operon observed in wildtype *E. coli* can be considered as a control system that ensures optimal growth of the organism. Under the selective pressure of evolution, the system might have evolved the *lac* operon to preferentially metabolize glucose in mixtures of sugars as it guaranteed an evolutionary advantage (faster growth) compared to substrate co-utilization (103).

As a new approach, dETFL avoids the pitfalls of simplifying modeling assumptions used in the current state-of-the-art computational models of metabolism and gene expression. Because of this, dETFL is the first dynamic ME-model formulation that can model lag phase and gradual proteome reconfiguration. However, despite these innovative findings, there are still drawbacks to dynamic constraint-based models that need refinement. For example, finding a good representative solution at each time step is extremely important. Here, we used the Chebyshev ball approach, as it is a single linear problem that is computationally simpler than other methods such as variability analysis or sampling. While we have reduced the computational burden of ME-models enough to efficiently perform iterative solving, there are new opportunities to further alleviate the computational cost of simulations. Directions to explore include fixing the integer variables of subproblems to reduce the NP-hardness of the model, and using quadratic programming, for instance, to perform an ellipsoid approximation of the enzyme solution space. Additionally, systematically reduced models, where less important parts of metabolic machinery are omitted, can also be used to reduce the complexity of the simulations (117, 118). With a reduced computational cost of simulations, exciting new research targets are also within reach, such as the dynamic effects of gene knock-outs or drug-induced changes in cell physiology.

The new computational formulations developed herein also offer new opportunities to test other hypotheses that explain diauxie. Succurro *et al.* (113) postulated the existence of two subpopulations of *E. coli*, where one obligately consumes glucose, while the other consumes acetate. Although the study of communities including thermodynamics-enabled ME-models is, for now, a computational challenge, cross-testing the hypothesis we present in this paper with a similar community-based context would certainly yield important insights on the respective role of proteome limitation and substrate competition in the emergence of diauxic behavior.

The inhibitory effect of glucose on certain parts of the metabolism is multiple, including catabolite repression, transient repression and inducer exclusion (119). Moreover, more complex regulation mechanisms are found in natural environments. For example, it has been shown that, on its natural marine substrate, the bacterium *Pseudoalteromonas haloplanktis* evolved regulation mechanisms allowing simultaneous diauxie and substrate co-utilization (120). Such high-order behavior might also have its origin in an optimal growth

Chapter 2. Dynamic ME-models: Emergence of diauxie as an optimal growth strategy under resource allocation constraints in cellular metabolism

program, and finding the biochemical constraints responsible for it would yield valuable insight on the optimal growth of organisms on complex media. In general, elucidating the emergence of regulation mechanisms in the context of evolutionary pressure will considerably increase our understanding and ability to engineer regulation systems, which are ubiquitous in biology, from wild-type *E. coli* to cancer cells. dETFL is an important step forward in this direction. Its use to uncover the optimality principles guiding the emergence of cellular regulatory control systems is key to a better understanding and, ultimately, mastery of metabolic engineering, be it applied to industrial hosts or the development of cell-based therapies.

2.4 Material and Methods

2.4.1 Rate of change of fluxes

One of the important points in the original formulation of dFBA is that the rate at which intracellular fluxes change is constrained. In the dFBA formulation, one imposed constraint is:

$$v(x, t + \Delta t) - v(x, t) \leq \dot{v}^{\max} \cdot \Delta t, \quad (2.7)$$

where $\dot{v}^{\max} \cdot \Delta t$ is defined as the maximum change of flux between two time points. However, the relationship between flux and enzyme concentration, as well as the dynamic mass balance, can be expressed in the following way (43):

$$v(x, t) \leq k_{\text{cat}} E, \quad (2.8)$$

$$\frac{dE}{dt} = v^{\text{syn}} - v^{\text{deg}} - v^{\text{dil}}, \quad (2.9)$$

with all the rates strictly positive. From this, it directly follows that

$$\dot{v}^{\max} = k_{\text{cat}} \dot{E}^{\max}, \quad (2.10)$$

$$-v^{\text{deg}} - v^{\text{dil}} \leq \dot{E}^{\max} \leq v^{\text{syn}}, \quad (2.11)$$

where we can rewrite \dot{E}^{\max} in a two components, one strictly positive, and the other strictly negative: $\dot{E}^{\max} = \dot{E}_+^{\max} - \dot{E}_-^{\max}$. Using expression relationships from ETFL, it is hence possible to bound the maximal rate change of fluxes in a fashion that is compatible with linear programming:

$$\dot{E}_{-}^{\max} \leq v^{\text{deg}} + v^{\text{dil}}, \quad (2.12)$$

$$\dot{E}_{+}^{\max} \leq v^{\text{syn}}. \quad (2.13)$$

These two constraints represent, respectively, the limitation in the decrease (dilution and degradation) and increase (synthesis) of the enzyme concentration. We can rewrite these in terms of dETFL variables:

$$0 \leq E_j^{t_{i+1}} - E_j^{t_i} \leq v_j^{\text{syn}} \cdot \Delta t, \quad (dEP_j)$$

$$0 \leq E_j^{t_i} - E_j^{t_{i+1}} \leq (v_j^{\text{deg}} + v_j^{\text{dil}}) \cdot \Delta t. \quad (dEN_j)$$

2.4.2 Variability in the estimation of macromolecule concentrations

A key element in ETFL is that macromolecule concentrations are an explicit variable in the optimization problem. In dETFL, these concentrations are important because they will constraint the feasible space for the calculation of next time step.

The formulation of ETFL relies on the approximation of the growth rate of the organism by a piecewise-constant function in the dilution term of the mass balances of macromolecules. This in turn allows the linearization of the bilinear term in the mass balances. However, this approximation has an error, which is given by the resolution η of the discretization. Given $\bar{\mu}$ the maximum growth rate of the model, and N the number of discretization points, the resolution of ETFL is given by $\eta = \frac{\bar{\mu}}{N}$. We can easily obtain the resolution of the estimation of a macromolecule concentration from this quantity.

The mass balance of a macromolecule X at concentration $[X]$ under steady state assumption is written in ETFL:

$$\frac{d[X]}{dt} = v^{\text{syn}} - v^{\text{deg}} - v^{\text{dil}}, \quad (2.14)$$

$$= v^{\text{syn}} - k_{\text{deg}} \cdot [X] - \mu \cdot [X], \quad (2.15)$$

$$= 0. \quad (2.16)$$

where v^{syn} , v^{deg} and v^{dil} are respectively the synthesis, degradation and dilution rates of the macromolecule, μ is the growth rate, and k_{deg} is the degradation rate constant of the macromolecule. In ETFL, μ is approximated by $\hat{\mu} = p\eta$, with $p \in \{0..N\}$. η is the

Chapter 2. Dynamic ME-models: Emergence of diauxie as an optimal growth strategy under resource allocation constraints in cellular metabolism

resolution of this approximation, which means, at all times:

$$\mu \in \left[\hat{\mu} - \frac{\eta}{2}, \hat{\mu} - \frac{\eta}{2} \right] \quad (2.17)$$

From Eq.2.15, and the relationship given in (2.17), we can rewrite:

$$[X] = \frac{v^{\text{syn}}}{k_{\text{deg}} + \mu} \quad (2.18)$$

$$\frac{v^{\text{syn}}}{k_{\text{deg}} + \hat{\mu} + \frac{\eta}{2}} \leq [X] \leq \frac{v^{\text{syn}}}{k_{\text{deg}} + \hat{\mu} - \frac{\eta}{2}} \quad (2.19)$$

We use this expression to represent the incertitude on the macromolecules concentrations at the previous time step, which is then used to constraint the current time step.

2.4.3 Backwards Euler integration scheme

At each time step, we operate an integration of the model between two time points. To this effect, using a robust integration scheme is necessary to guarantee a solution quality that is as good as possible. We chose to use a backwards (implicit) Euler integration scheme given its ability to handle stiff problems (121). Usually, a drawback of implicit schemes is that they require to solve an implicit equation to define the state of the system at each time step. In contrast, explicit methods simply require to apply a defined set of calculations (*e.g.* a linearized state function) on the current state. In our case, however, there is little cost associated to using an implicit method rather than the explicit Euler method, since we already need to solve a whole MILP problem to compute the solution to the dETFL problem at each time step.

In this context, we can rewrite Eqs. dEP_j and dEN_j in their Euler-form:

$$0 \leq E_j(t_{i+1}) - E_j(t_i) \leq v_j^{\text{syn}}(t_{i+1}) \cdot \Delta t, \quad (dEP_j)$$

$$0 \leq E_j(t_i) - E_j(t_{i+1}) \leq \left(v_j^{\text{deg}}(t_{i+1}) + v_j^{\text{dil}}(t_{i+1}) \right) \cdot \Delta t, \quad (dEN_j)$$

where E_j is the concentration of a given enzyme at the previous time step, $E_j(t_{i+1})$, $v_j^{\text{deg}}(t_{i+1})$, $v_j^{\text{deg}}(t_{i+1})$, $v_j^{\text{dil}}(t_{i+1})$ are variables of the dETFL problem in the next time step. $E_j(t_i)$ is a variable constrained around the value of the previous solution, as explained in the previous section.

2.4.4 Chebyshev center

One important issue when dealing with both (mixed-integer) linear optimization and iterative solving is the multiplicity of solutions. Indeed, the optimality principle in LP only guarantees a unique global optimum value for the objective, but not a unique optimal solution for the variables. In fact, at each time point, there is most often a (piecewise-)continuum of solutions (including flux values, macromolecule concentrations ...) that can satisfy a maximal growth rate, while describing different physiologies. For example, two optimal states, using different pathways with a similar enzyme cost, will yield different proteomes and associated fluxes. In addition, due to the constraints applied on the rate of change of macromolecule concentrations, in each subsequent time point, the proteome, transcriptome and flux values will be dependent on all the previous solutions. Because of these two factors, each new realisation of the integration procedure might yield different results.

An additional issue is that simplex-based solvers tend to give sparse and extremal results (corners of the explored simplex), which do not represent accurately the full extent of the considered solution space. Several methods can alleviate these issues, all based on finding a good representative of the solution space. One first solution is to use as observation the mean of the variability analysis, rather than a single optimal solution. This however requires $\mathcal{O}(2n)$ optimizations to be carried out. Another way would be to sample the feasible space, but the sheer size of dETFL models makes sampling impractical. The Supplementary Figure S7 shows a 2-D example of the difference between these 3 approaches. The method we chose is to try to find the maximally inscribed sphere in the solution space. The center of this sphere, called the Chebyshev center, can be found by optimising a single linear problem if the solution space is a polytope(122). It is the case in dETFL, as the problem is defined with linear inequality constraints.

In the case of a polyhedron defined by inequalities of the form $a_i^\top x \leq b_i, x \in \mathbb{R}_+^n$, finding the Chebyshev center of the solution space amounts to solving the following optimization problem:

$$\begin{aligned} & \underset{r, x}{\text{maximize}} && r \\ & \text{subject to} && a_i^\top x + \|a_i\|_2 r \leq b_i \end{aligned} \tag{2.20}$$

This is similar to adding a common slack to all inequalities and maximize its size, which maximizes the distance of the solution to the inequality constraints. However, not all variables and constraints need to be considered in the definition and inscribing of this sphere. In particular, we are interested in a representative solution for macromolecule concentrations, which only play a role in a limited set of constraints. To this effect, we define \mathcal{I}_c and \mathcal{J}_c , respectively the set of inequality constraints and variables with respect to which the Chebyshev center will be calculated. Let us also denote \mathcal{E} the set of equality

Chapter 2. Dynamic ME-models: Emergence of diauxie as an optimal growth strategy under resource allocation constraints in cellular metabolism

constraints of the problem, a_i, c_i respectively the left-hand side of the inequality and equality constraints, and b_i, d_i their respective right-hand side. From there, we can define the modified centering problem:

$$\begin{aligned}
& \underset{r, x}{\text{maximize}} && r \\
& \text{subject to} && \mu = \mu^*, \\
& && a_i^\top x + \|\mathbb{1}_{\mathcal{J}_c} \circ a_i\|_2 r \leq b_i, \quad \forall i \in \mathcal{I}_c, \\
& && a_i^\top x \leq b_i, \quad \forall i \notin \mathcal{I}_c, \\
& && c_k^\top x = d_k, \quad \forall k \in \mathcal{E},
\end{aligned} \tag{2.21}$$

where μ^* is the maximal growth rate calculated at this time step, r the radius of the Chebyshev ball, x the column vector of all the other variables of the ETFL problem, $\mathbb{1}_{\mathcal{J}_c}$ has for j^{th} element 0 if $j \in \mathcal{J}_c$, else 1, and \circ denotes the element-wise product between two vectors. Thus, $\|\mathbb{1}_{\mathcal{J}_c} \circ a_i\|_2$ is the norm of the projection of the constraint vector onto \mathcal{J}_c . We show an example illustration in 3D in the Supplementary Figure S8.

For enzymes, for example, it is akin to making the model produce more enzymes than necessary to carry the fluxes, while respecting the total proteome constraint. By maximizing the radius of the sphere inscribed in the solution space, at maximal growth rate, we are effectively choosing a representative solution of the maximal growth rate feasible space. We then use this solution as a reference point for the next computation step.

All simulations in this paper perform Chebyshev centering on enzyme variables at each time step.

2.4.5 Initial conditions

Since dETFL is an iterative method, it is necessary to set an initial reference point (initial conditions) from which the dynamic analysis will integrate over time. The initial solution is set up as follows:

1. Set typical uptake fluxes for carbon sources and oxygen,
2. Perform a growth maximization using ETFL
3. Fix the growth to the optimum,
4. Find the Chebyshev center of the solution space.

The solution reported by the latter optimization problem is then used as a starting solution for the dETFL analysis.

2.4.6 Extracellular concentrations

At each time step, extracellular concentration are updated following a standard Euler scheme, similarly to what is done in Mahadevan *et al.* (64). The extracellular concentrations of glucose, lactose, and acetate, follow a system of ordinary differential equations:

$$\frac{d[Glc]}{dt} = v_{glc} \cdot X, \quad (2.22)$$

$$\frac{d[Ac]}{dt} = v_{ac} \cdot X, \quad (2.23)$$

$$\frac{d[Lcts]}{dt} = v_{lcts} \cdot X. \quad (2.24)$$

We linearize this system into the following forward Euler scheme:

$$[Glc]_{t+1} = [Glc]_t + v_{glc} \cdot X \cdot \Delta t, \quad (2.25)$$

$$[Ac]_{t+1} = [Ac]_t + v_{ac} \cdot X \cdot \Delta t, \quad (2.26)$$

$$[Lcts]_{t+1} = [Lcts]_t + v_{lcts} \cdot X \cdot \Delta t. \quad (2.27)$$

We use these linearized equations to update the extracellular medium after the solution to each time step has been computed.

2.4.7 Model

The model used is the the vETFL model of iJO1366, presented in the original ETFL publication (43). 15 additional enzymes were added to the model to properly account for the protein cost of transporting glucose, lactose and galactose from periplasm to the cytoplasm. A simplified metabolic map of the glucose, lactose, and galactose pathways to G6P is shown in Fig. 2.3.

2.4.8 Kinetic information

The Michaelis-Menten parameter $K_M^{glc} = 0.015$ mM for glucose was taken from the original dFBA paper (64). The $K_M^{lcts} = 1.3$ mM of for lactose was obtained from a study by Olsen *et al.* on the specificity of lactose permeases (123). Details on the added enzymes are available in the Supplementary Table S2. The Michaelis-Menten parameter $V_{max}^{glc} = 15$ mM was used similarly to previous work (64).

Chapter 2. Dynamic ME-models: Emergence of diauxie as an optimal growth strategy under resource allocation constraints in cellular metabolism

Because of incertitude in the values found in the literature, V_{\max}^{lcts} was directly computed from the catalytic rate constants of enzymes consuming periplasmic lactose (LACZpp, LCTStpp, LCTS3ipp). Since ETFL gives access to enzyme concentrations, we can rewrite the expression of V_{\max}^{lcts} using catalytic rate constants k_{cat}^j :

$$V_{\text{lcts}}^{\max} = \sum_{j \in \mathcal{L}} k_{\text{cat}}^j \cdot [E_j], \quad (2.28)$$

where \mathcal{L} is the set of periplasmic enzymes consuming lactose. Taking this into account allows to replace the parameter V_{\max}^{lcts} by an explicit internal variable.

Acetate transport is assumed to be mostly diffusive (124), and its secretion rate was bounded at $5 \text{ mmol.gDW}^{-1}.\text{h}^{-1}$, and its uptake to $3 \text{ mmol.gDW}^{-1}.\text{h}^{-1}$. These values are of the same order of magnitude as in previous studies (27, 64, 113) Oxygen gas-liquid transfer kinetic parameters are taken from Mahadevan *et al.* (64).

2.4.9 Implementation

The code has been implemented as a plug-in to pyTFA (94), a Python implementation of the TFA method, and ETFL (43), an implementation of ME-Models accounting for expression, resource allocation, and thermodynamics. It uses COBRApy (95) and Optlang (96) as a backend to ensure compatibility with several open-source (GLPK, scipy) as well as commercial (CPLEX, Gurobi) solvers. The code is freely available under the APACHE 2.0 license at <https://github.com/EPFL-LCSB/etfl>.

2.5 Supporting Information Appendix (SI)

Supplementary Note S1

Note on the DynamicME Assumptions

Supplementary Table S2

Properties of glucose and lactose transporting reactions and enzymes. Reaction names from the original iJO1366 model (48). Enzyme symbols adapted from Biocyc (65). k_{cat} values taken from Lloyd *et al.* (41).

Supplementary Figure S3

Enzyme levels of the glucose pathway, in the glucose/lactose diauxie experiment with glucose pre-culture.

Supplementary Figure S4

Enzyme levels of the lactose pathway, in the glucose/lactose diauxie experiment with glucose pre-culture.

Supplementary Figure S5

Enzyme levels of the glucose pathway, in the glucose/lactose diauxie experiment with lactose pre-culture.

Supplementary Figure S6

Enzyme levels of the lactose pathway, in the glucose/lactose diauxie experiment with lactose pre-culture.

Supplementary Figure S7

2D example the different schemes to find a representative point of the space: variation analysis, sampling, or Chebyshev centering.

Supplementary Figure S8

3-Dimensional example of a Chebyshev center. The feasible space is denoted by the polytope \mathcal{C} . The Chebyshev center with respect to variables E_1 and E_2 is X_{E_1, E_2} . It is the center of the largest 2-D sphere on a plane parallel to (E_1, E_2) that is inscribed in \mathcal{C} . This sphere exists on the plane \mathcal{P} , materialized in light blue.

Supplementary Figure S9

Enzyme composition of the conceptual model when no constraints are applied to the rate-of-change of enzyme concentrations.

Supplementary Figure S10

Figure for the switched k_{cat} experiments.

Acknowledgements


The authors would like to thank Dr. Ljubiša Mišković and Dr. María Masid Barcón for their valuable input on this manuscript. Pierre Salvy would like to express his gratitude to Kilian Schindler and Prof. Daniel Kuhn for the valuable discussion around the formulation of the Chebyshev problem. This work has received funding from the European Union's Horizon 2020 Research and Innovation Programme under the Marie Skłodowska-Curie grant agreement No 722287, the European Union's Horizon 2020 Research and Innovation Programme under grant agreement No 686070, and the Ecole Polytechnique Fédérale de Lausanne (EPFL).

Author Contributions

P.S. and V.H. designed the study. P.S. wrote the code, curated the data, and performed the studies. P.S. and V.H. wrote the manuscript.

Competing Interests

The authors declare no competing interests.



3. Emergence of the Crabtree effect in a model of metabolism and expression for *S. cerevisiae*

Chapter 3. Emergence of the Crabtree effect in a model of metabolism and expression for *S. cerevisiae*

Omid Oftadeh¹, Pierre Salvy¹, Maxime Curvat¹, María Masid Barcón¹, Ljubiša Mišković¹, Vassily Hatzimanikatis^{1,*},

¹ Laboratory of Computational Systems Biotechnology, École Polytechnique Fédérale de Lausanne (EPFL), Lausanne, Switzerland

* Corresponding author: vassily.hatzimanikatis@epfl.ch

This chapter details an important milestone in models of metabolism and expression. We describe the first ME-model of a eukaryotic organism, *S. cerevisiae*, also known as baker's yeast. An important mechanism of *S. cerevisiae* physiology is its ability to perform ethanol fermentation while being in an oxygenated environment, and in presence of an abundance of glucose, also known as the Crabtree effect. Such behavior seems at odds with the idea of optimal growth in stoichiometric models, because fermentation does not appear necessary to the system's metabolism, and yet each carbon lost as fermented ethanol is carbon that does not participate in the biomass. We show in this chapter that the Crabtree effects stems from the constraints on the proteome of *S. cerevisiae*, and that our models quantitatively captures this effect without fitting experimental data.

The chapter is adapted from a manuscript in preparation. Vassily Hatzimanikatis, Omid Oftadeh, and I designed the study and wrote the article. Omid and I wrote the code to adapt ETFL to eukaryotic organisms. Omid Oftadeh ran the simulations and did the enzymatic data curation. I provided Omid with methods and scripts I used to write ETFL for *E. coli*, and helped him adapt them to his use case. Maxime Curvat, under the supervision of María Masid Barcón and Ljubiša Mišković, curated the thermodynamics data for the yeast model. María Masid Barcón performed extensive quality control of the annotations, and both María Masid Barcón and Ljubiša Mišković wrote the section on thermodynamic curation.

All the code and documentation is available under the APACHE 2 license at:

<https://github.com/EPFL-LCSB/yetfl>

<https://gitlab.com/EPFL-LCSB/yetfl>

3.1 Introduction

Metabolic networks are the most widely studied and modeled type of biological networks, with more than six thousand genome-scale metabolic models (GEMs) being reconstructed for archaea, bacteria, and eukaryotes (24, 34). The GEMs are created by associating the genes in an organism’s genome with enzymes and reactions in databases, which in turn, enables the study of the the organism’s phenotype based on its genotypic information.

In particular, Flux Balance Analysis (FBA) is a constraint-based optimization technique that allows to compute the metabolic flux of each reaction in a metabolic network by formulating a linear programming problem and optimizing an objective function to select a solution in the feasible space (28). The objective function is typically chosen to represent biologically relevant objectives such as selection pressure (31, 32). Despite its wide applicability, FBA is unable to predict some important features of metabolic networks. Indeed, it has been found to predict biologically irrelevant solutions, including cycles with unrealistically high fluxes (53), or thermodynamically infeasible solutions (35, 36). Moreover, FBA is unable to account for limited catalytic capacity of the enzymes and does not model the cellular expression system.

To improve FBA and its modeling abilities, additional constraints that represent biological phenomena, either from empirical or mechanistic evidence, have been introduced. Thermodynamic flux analysis (TFA) (35, 36) was developed along this idea, and it enforces the coupling of reaction directionality with the reactions’s Gibbs free energy to produce thermodynamically feasible solutions. More importantly, it also enables metabolomics data integration through the addition of variables representing metabolite concentrations.

The GECKO formulation (Genome-scale models with Enzymatic Constraints using Kinetic and Omics data) adapts FBA to account for limited catalytic activity of enzymes by including enzyme concentrations as variables to the constraint-based problem (38). The method is able to capture a realistic maximum specific growth rate, proteome-limited growth, and the occurrence of overflow metabolism, in *S. cerevisiae* (38). The introduction of explicit enzyme concentrations as variables also enables the direct integration of proteomics data. However, GECKO does not explicitly consider the cost of protein synthesis. Instead, it assumes that a peptide’s share in a protein pool is proportional to the inverse of its molecular weight. In this context, the molecular weight represents the cost of the enzyme in terms of proteome allocation. However, the actual cost of synthesizing the enzyme is absent from the formulation. Therefore, the method does not account for the competition for amino acids, energy, and polymerizing enzymes between different proteins, while all the latter phenomena play an important part in the expression of genes.

Metabolic and Expression models (ME-models) form another class of constraint-based models, which in addition to the metabolic and the catalytic constraints include the

cellular expression system (39, 40, 43). ME-models can predict the concentration of each mRNA and enzyme while considering its cost of synthesis and a total cellular expression capacity. The original formulation of these models allows the integration of proteomics and transcriptomics data. However, this same original formulation of ME-models was incompatible with the integration of thermodynamics constraints, and consequently metabolomics data could not be incorporated. The presence of bilinear terms in the original formulation of ME-models warranted the use of special nonlinear optimization procedures and high precision solvers (45, 46, 41). As a result, adding thermodynamic constraints into ME-model formulations would necessitate solving a Mixed-Integer Nonlinear Programming (MINP) problem. Indeed, thermodynamic constraints from TFA require integer variables, making TFA a Mixed-Integer linear Program (MILP).

To address these issues, a new formulation for ME-models was recently proposed. The approach, called Expression and Thermodynamics-enabled Flux models (ETFL)(43), avoids bilinear terms by discretizing growth and solving locally linearized mixed-integer problems instead of a MINP problem. Like previously published ME-models (40, 41), the ETFL model was developed for *E. coli*. However, the ETFL formulation can readily be extended to study eukaryotic organisms.

S. cerevisiae is one of the industrially most relevant organisms (125, 126), and it has been widely used for biological and medical research studies (127). Because of its ubiquity in metabolic engineering, several GEMs of this organism have been published over the years (128, 129, 130, 131, 132, 133). Despite its industrial significance and the academic interest around this microbe, so far no ME-model of *S. cerevisiae* has been developed. This might be partly because eukaryotic organisms require additional considerations in modeling the compartmentalized cellular expression system. Here, we propose a ME-model for *S. cerevisiae*, yETFL, which is based on an extended ETFL formulation. The model also includes a thermodynamic curation of its metabolites and reactions, and can readily integrate metabolomic, proteomic, and transcriptomic data. The presented methodological developments of ETFL also pave way for a generalized development of ME-models for other eukaryotes.

3.2 Results and Discussion

3.2.1 Model description

Four ME-models of *S. cerevisiae* were constructed from the consensus Yeast8 model published by Lu *et al.* (133). The models boundary fluxes are set to correspond to a minimal medium, with only inorganic metabolites and a carbon source. A general description of different yETFL models and their features is provided in Table 3.1, and the nomenclature of the different models used is detailed in Table 3.2.

Table 3.1. Properties of the vETFL model generated from Yeast8.

Growth upper bound $\hat{\mu}$	$0.75h^{-1}$
Number of bins N	128
Resolution $\frac{\hat{\mu}}{N}$	$0.0058h^{-1}$
Number of constraints	92429
Number of variables	66746
Number of species	
– Metabolites	2689
– Peptides	1393
– mRNAs	1393
– rRNA	6
Number of enzymes	
– Metabolic	1059
– RNA polymerases	2
– Ribosomes	3
Number of reactions	
– Metabolic	2678
– Transport	1047
– Exchange flux	243
– Transcription	1393
– Translation	1393
– Complexation	1065
– Degradation	2458
Number of metabolites $\Delta_f G'^o$	2433
Number of reactions $\Delta_r G'^o$	3184
Percent of metabolites $\Delta_f G'^o$	90%
Percent of reactions $\Delta_r G'^o$	80%

yETFL comprises 1059 proteins coupled to 2588 reactions. Among these, we found the catalytic rate constants (k_{cat}) for 943 enzymes. The catalytic rate constant of 39 enzymes were approximated by the median k_{cat} value in *S. cerevisiae*, and for 77 enzymes that are associated with 167 transports we assigned arbitrarily high k_{cat} value, to ensure that the corresponding reactions are not catalytically constrained (see Materials and Methods). Among all the proteins, there exist 107 complexes; the others are monomeric enzymes composed of a single peptide.

3.2.2 Prediction of specific growth rate

A traditional test to assess the quality of a genome-scale model of metabolism is to predict the maximal specific growth rate of the organism it models. However, FBA predicts that the growth rate increases linearly with carbon uptake. In reality, however, the growth rate reaches a plateau at high substrate uptake. Thus, the range in which FBA predicts correct growth rates is limited. In particular, the finite proteome size of an organism is known

Chapter 3. Emergence of the Crabtree effect in a model of metabolism and expression for *S. cerevisiae*

Table 3.2. Nomenclature of the models used in the study of the *S. cerevisiae* model Yeast8. EFL stands for Expression and Fluxes, ETFL for Expression, Thermodynamics, and Fluxes. The c- and v- prefixes indicates the inclusion of constant or growth-dependent biomass compositions

	growth-independent biomass composition	growth-dependent biomass composition
(-) thermodynamics	cEFL	vEFL
(+) thermodynamics	cETFL	vETFL

for being a limiting factor in both growth rate and carbon uptake (37, 101, 134). Since ETFL, and ME-models in general, account for gene expression capacity, including the limited proteome and transcriptome sizes, it is able to predict proteome-limited growth (39, 40, 43).

ETFL can model proteome (and transcriptome) limitation in several ways. The biomass composition is directly tied to the size of the proteome and of the transcriptome, since these two will account for mRNA and protein content in the cell. The study of models with different biomass compositions thus entails the study of models with different proteome and transcriptome sizes. Thus, we investigated the usage of both constant (cETFL) and growth-dependent (vETFL) biomass compositions by simulating the maximal growth rates predicted by our thermodynamically enabled ME-models at various glucose uptakes (Fig. 3.1). In both cases and in contrast to FBA, the growth rate reached a plateau at higher values of the glucose uptake rate, which is in accordance with the experimental results by Van Hoek *et al.* (135). When glucose is in excess, higher fluxes require higher amounts of enzymes. However, the limitations in the expression system and the catalytic activity of enzymes prevent the growth rate from increasing further. High growth rates also incur high dilution rates, which further increases the cost of carrying higher fluxes. In particular, we observe a shift from glucose-limited growth to proteome-limited growth at uptakes around $4 - 5 \text{ mmol} \cdot \text{g}_{\text{DW}}^{-1} \cdot \text{h}^{-1}$. The maximum predicted growth rate with cETFL was 0.46 h^{-1} , while vETFL predicted 0.42 h^{-1} . These results agree with the experimentally measured maximal specific growth rate in the literature that are commonly in the range of $0.4\text{-}0.5 \text{ h}^{-1}$ for different strains in a minimal medium. (136, 137). We also observe that the vETFL model grows faster than the cETFL model at low uptake rates; the trend is reversed at higher growth rates. Since cETFL is extracted from the FBA biomass composition, which is fixed at an average point, it is indeed expected that a model with a better granularity captures lower requirements at low growth rate (increased biomass yield, and hence, slope), and higher ones at high growth rate (decreased biomass yield, and hence, slope).

We observe small discrepancies in the maximal growth rate between the experimental data and the yETFL results for the glucose uptake rate ranging from $\approx 4 \text{ mmol} \cdot \text{g}_{\text{DW}}^{-1} \cdot \text{h}^{-1}$ to $\approx 12 \text{ mmol} \cdot \text{g}_{\text{DW}}^{-1} \cdot \text{h}^{-1}$. A possible cause for these discrepancies might be the growth-dependence of some parameters, such as the ribosomal elongation rate. To avoid over-constraining the model, we used the highest reported values for ribosomal elongation

rate, which usually occurs at the high growth rates (47, 138). Since our formulation accounts for growth-dependent parameters, as soon as our knowledge about the variation of these factors with growth rate is augmented, we can integrate this knowledge into the formulation of yETFL. Another explanation of the difference might come from the regulation system used by *S. cerevisiae* when transitioning from nutrient-limited growth to proteome-limited growth. It is important to remember ME-models work under the assumption of optimality (in this case, maximal growth rate), and that the cellular system evolved under selection pressure to match this optimality. In this context, the regulatory network of *S. cerevisiae* can be seen as a control system that pushes the metabolism towards optimality, and deviations from model optimality in transition regions are simply limitations of the regulatory system¹⁵.

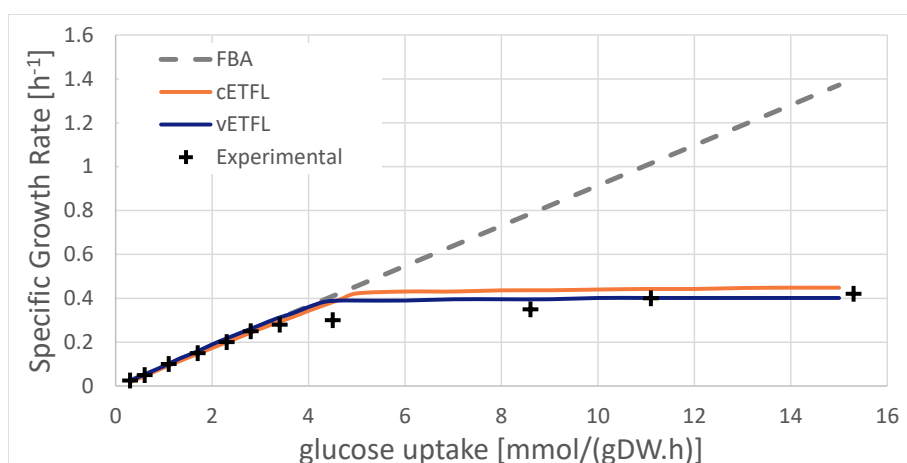


Fig. 3.1. Growth rate of the yETFL models (orange and blue) and Yeast8 FBA (dashes) with respect to glucose uptake. Experimental data (crosses) by Van Hoek *et al.* (135)

3.2.3 Gene essentiality analysis

A classic way to ascertain the accuracy of a GEM is to compare *in silico* gene knock-outs with *in vivo* data from the Stanford Yeast Deletion datasets. In gene essentiality analysis, we simulate the growth of an organism for each gene knock-out and compare the predicted and experimental values using a confusion matrix, which describes true and false positives (the cell grows without the gene) or negative (the cell needs the gene to survive). The gene essentiality results for the metabolic genes are slightly improved in cEFL compared to the FBA models (Fig. 3.2-b and -c). This shows that we were able to conserve the quality of gene-reaction associations moving from the FBA to the ETFL formalism. Compared to the FBA model, ETFL models include genes that correspond to RNA polymerases and ribosomes, which is also why the amount of true negatives (the gene is essential

¹⁵This interpretation is an important topic of chapter 2

and predicted as essential) is increased. The cETFL model predicted five essential genes that were predicted as non-essential using FBA (Fig. 3.2-d). The increased number of essential genes in the cETFL is due to the more restricted reaction directionalities after the integration of thermodynamic constraints.

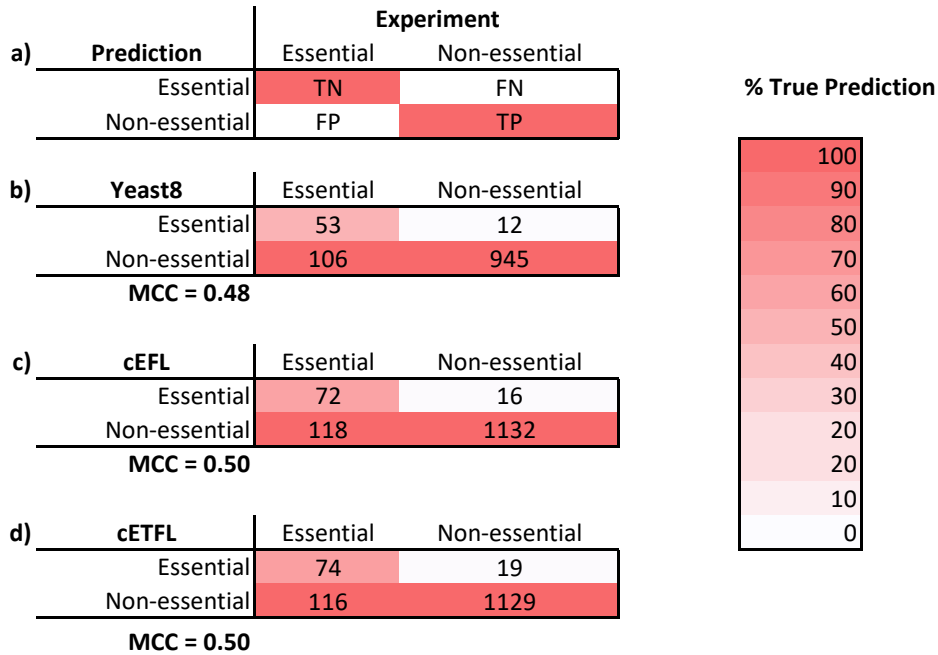


Fig. 3.2. Confusion matrices for gene essentiality studies. **a.** Conventions from used for gene essentiality. TN is True Negative. FN is False Negative. FP is False Positive. TP is True Positive. The color shading represents how good the classification is in the experimental class. Perfect classification should have a strict red first diagonal, as shown on this example. **b.** Gene essentiality prediction for the FBA model Yeast8, yielding a Matthew's correlation coefficient (MCC) of 0.48. **c.** Gene essentiality prediction for the genes expressed in the cEFL model of Yeast8, yielding a MCC of 0.50. **d.** Gene essentiality prediction for the cETFL model of Yeast8, yielding a MCC of 0.50.

3.2.4 Crabtree effect

Overflow metabolism a phenomenon in which organisms produce energetic carbon compounds as byproduct of their growth in the presence of excess carbon substrate (135, 139, 140) (98). Such behavior can seem surprising, as these carbon compounds could theoretically still be used for biosynthesis, and their production directly decreases the yield of an organism's biomass synthesis – a seemingly competitive disadvantage. Several hypotheses can justify overflow metabolism, a chief one among them being the limited proteome capacity of a cell.(37, 101, 134)¹⁶. Overflow metabolism in yeast is called the Crabtree effect: after a critical growth rate, which is strain-specific but usually

¹⁶This topic is also discussed in chapter 2.

close to 0.3 h^{-1} , the cells shift from pure respiration to a combination of respiration and fermentation, in the presence of oxygen excess. Fermentation is an energy-producing process that is less efficient than the aerobic respiration, but does not depend on the presence of molecular oxygen. It also produces byproducts, under the form of CO_2 and ethanol in *S. cerevisiae*.

FBA is not able to predict the Crabtree effect unless some ad hoc changes are made in the constraints or the objective function (101). Since yETFL considers both the limitations in the catalytic capacity of the enzymes and the protein expression machinery, it is able to predict the metabolic shift of overflow metabolism at higher growth rates (Fig 3.3). We observe the simulation predicts at growth rates higher than $0.38 - 0.40 \text{ mmol} \cdot \text{g}_{\text{DW}}^{-1} \cdot \text{h}^{-1}$ the secretion of ethanol, and an increase in CO_2 production, while O_2 consumption is reduced. The model showed good qualitative agreement with the experimental data from Van Hoek *et al.* (135) in aerobic, glucose-limited chemostat fermentation. The vE(T)FL models present an onset of Crabtree effect that is earlier than the one for the cE(T)FL models (Fig 3.3). Since the vETFL model has a different (growth-dependent) proteome capacity from that of cETFL (constant), it is expected that overflow metabolism does not occur at the same glucose uptake.

Experimental data suggests the Crabtree effect should happen earlier than what the models predict. We believe the late predictions of our models stem from the conservative (higher) values chosen for the catalytic rate constants of enzymes (see Materials and Methods). As such, less enzymes are needed to carry more flux, which pushes the point at which the proteome capacity becomes limiting.

It is worth noting that yETFL was able to capture the occurrence of the Crabtree effect only by integrating experimentally measured parameters and without making ad hoc modifications in the model or in the formulation. In particular, no proteomics were used to show the onset of the effect, which stems purely from the stoichiometric and expression constraints.

3.3 Conclusion

In this work, we developed a ME-model for a eukaryotic organism, *S. cerevisiae*. The adaptation of the ETFL formulation required to consider compartmentalized expression systems, with separate ribosomes and RNA polymerases. We validated the growth predictions of the model against experimental data, and showed its gene essentiality predictions are on par with the state-of-the-art FBA model of yeast. With a model that represents and constrains the proteome of the cell, we were able to reproduce emergence of the Crabtree effect, and observe the secretion of ethanol in aerobic conditions — without actually integrating experimental data, as opposed to previous descriptions of the Crabtree effect (38).

Chapter 3. Emergence of the Crabtree effect in a model of metabolism and expression for *S. cerevisiae*

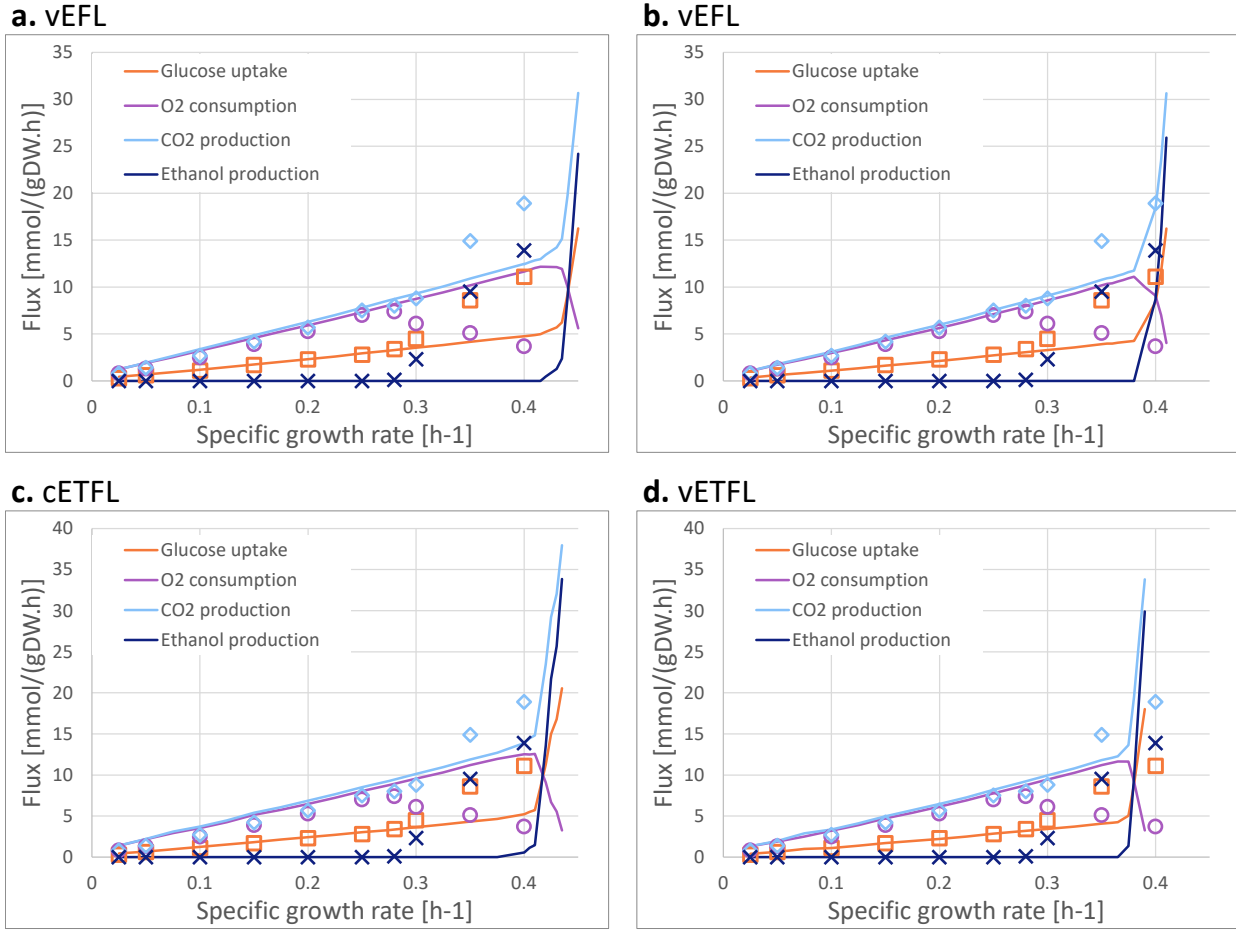


Fig. 3.3. Comparison of experimental data from Van Hoek *et al.* (135) (crosses) and simulation results (lines) for the Crabtree effect. Absolute exchange rates of glucose (orange), O₂ (purple), CO₂ (light blue), and ethanol (dark blue) are shown. **a.** cEFL model. **b.** vEFL model. **c.** cETFL model. **d.** vETFL model.

ME-models improve upon models of metabolism by considering the coupling between metabolism and expression system. Such detailed description of the gene expression mechanisms allows the elucidation of intracellular behaviors underlying cell physiology, for instance the Crabtree effect in this work. Another advantage of ETFL is it allows a direct extension of the model to other types of analyses. We studied here the Crabtree effect in a steady state manner; yet understanding its emergence in a dynamic setting, as done previously for the *E. coli* overflow metabolism (98) will yield valuable insights on the optimality of the regulation mechanisms in yeast. Additionally, the ETFL formulation also allows the integration of multiple types of data, such as proteomics, transcriptomics, metabolomics, and fluxomics. Integrating experimental measurements from strains of interest will also help better characterize their intracellular state, a crucial element to

better understand and engineer cell lines.

The method presented here is readily adaptable to any eukaryotic organism for which a well-curated GEM is available. The quality of the information about enzymes (catalytic rate constants, composition) will change the quantitative predictions of the model, but average or estimated values still provide an acceptable replacement, as long as key enzymes such as ribosomes, polymerases, and ATP synthase are properly represented (43). ETFL-based ME-models can always be improved as more experimental data become available.

Eukaryotic organisms are extremely important in industrial biotechnology: *S. cerevisiae* is a popular host organism for the bioproduction of fuels, specialty and commodity chemicals (7, 8, 9, 10, 11), and Chinese hamster ovary (CHO) cells are the main platform organism used for the production of therapeutic proteins (141). The prevalence of these industrial organisms drives a need for better models of the metabolism and gene expression. In this context, the availability of eukaryotic ME-models will help improve the understanding and engineering of industrial hosts, helping create ever more efficient and productive cell lines.

Last, but not least, humans are also eukaryotes. Eukaryotic ME-models open the way to models of metabolism and expression being used in health and medicine. Metabolic models are already being used to elucidate the intracellular state of immune cells in response to inflammation (142). ME-models have the capacity to integrate even more data to represent tissue-specific models, and take into account their biochemical environment. Their versatility makes them ideal candidates to study complex cell types, such as immune cells at different stages, or cells in a tumor micro-environment.

3.4 Materials and Methods

3.4.1 Formulation of the ETFL model

yETFL is based on the ETFL formulation which was previously described in details in Salvy *et al.* (43). The ETFL constraints can be divided into five main categories:

- **Steady-state constraints:** Enforce all metabolite and macromolecule concentrations to be at steady-state. For metabolites, these constraints are the same as in FBA.
- **Thermodynamic constraints:** Couple the directionality of reactions with their Gibbs free energy. These constraints are the same as in TFA.
- **Catalytic constraints:** Define upper bounds on the reaction fluxes based on the enzymatic capacity of the associated enzymes.

- Expression constraints: Model the synthesis of mRNAs, peptides, and proteins, and constrain synthesis rates based on the limitations of transcription and translation machinery.
- Allocation constraints: Determine the available amounts of DNA, RNA, and proteins in the cell. The ETFL formulation allows modeling the growth-dependent abundance of these macromolecules, if experimental data is available. Whenever the experimentally measured abundance of these macromolecules during the growth is not available, we assume that the ratio between these quantities is growth-independent, an assumption also made in FBA.

Depending on whether we include thermodynamic constraints or not and the type of resource allocation (either constant or variable), we developed four different types of models (Table 3.2).

3.4.2 Ribosomes and RNA polymerases

To model the ribosomes and the RNA polymerases, information about their constituting peptides (and ribosomal RNA) and catalytic rate constants is required. The previous ME-models were constructed for bacteria (39, 40, 43), with one ribosome and one RNA polymerase being sufficient to represent the cellular expression machinery. In contrast, yeast being a eukaryotic organism, it features an additional mitochondrial ribosome and RNA polymerase. To consider this complexity, we defined multiple RNA polymerases and ribosomes in the model:

RNA Polymerase Similarly to the other eukaryotes, yeast has three different types of nuclear RNA polymerases. However, most of the mRNA transcripts are transcribed by RNA polymerase II (143, 144). In yETFL, we implemented this nuclear RNA polymerase, and we modeled that all the nuclear genes could be transcribed only by this enzyme, similar to the previous work (43). For mitochondrial genes, we defined a mitochondrial RNA polymerase characterized by its own composition and kinetic parameters (144). Similarly, the mitochondrial genes were only allowed to be transcribed by this polymerase.

Ribosomes The structure of the cytosolic ribosomes in yeast contains four ribosomal RNA (rRNA) molecules encoded by four different genes. In addition to these four rRNAs, the cytosolic ribosomes contains 78 peptides encoded by 137 genes (59 peptides are encoded by two alternative genes) (145). To account for these alternative ribosomal peptides, we defined two cytosolic ribosomes: one ribosome was constructed by the first set of peptides (designated with 'A' in their standard names) and the other one was constructed with the alternative genes (designated with 'B' in their standard names).

We assumed both cytosolic ribosomes had the same elongation rate. A mitochondrial ribosome was defined to translate mitochondrial genes. This ribosome is composed of by two rRNAs and 78 peptides (146).

Data Collection

Genome-scale metabolic model The most recent GEM of *S. cerevisiae*, Yeast8 (133), was used as a basis to construct our model. The following modifications to Yeast 8 were made:

- Pseudometabolites defined for RNAs and proteins as well as pseudoreactions defined for their synthesis were replaced by the explicit expressions for RNA and protein synthesis, according to the procedure described in the supplementary material of Salvy *et al.* (43).
- tRNAs and their reactions and were adapted into a formulation that accounts for dilution effects according to the ETFL procedure. This is necessary as tRNAs are species for which the dilution effect is not necessarily negligible.
- The biomass reaction was modified to account for growth-dependent composition, as discussed in the section Modifying the growth-associated maintenance.

The latest published version of Yeast8 model, Yeast8.3.4, was obtained from GitHub as it was provided by Laboratory of Systems and Synthetic Biology at Chalmers University (<https://github.com/SysBioChalmers/yeast-GEM>).

Thermodynamic curation Information about the thermodynamic properties of reactions allows us to integrate the available metabolomics and fluxomics data into models, and to compute thermodynamically consistent values of metabolic fluxes and metabolite concentrations that were not measured. We used the group contribution method (GCM) (147), to determine the standard Gibbs free energy of formation in aqueous, ionic environments(148) for 1092 out of 1326 (82%) unique metabolites from Yeast8. We were not able to determine the thermodynamic properties for remaining 234 (17.6%) metabolites, because: (i) 89 (6.7%) metabolites represented modeling species such as pools of proteins, nucleotides, lipid chains; and (ii) 145 (10.9%) metabolites were with unknown molecular structure or they contained structural groups for which the estimated standard Gibbs energy of formation is unknown (e.g., Acyl Carrier Protein group). Using the information about the standard Gibbs energy of formation of compounds, we estimated the standard Gibbs free energy of reactions for 3184 out of 3991 (80%) reactions from Yeast8.

mRNA, Peptide, and Protein data The sequences for the peptides and mRNAs were downloaded obtained from the KEGG database (149). Information about the stoichiometry of peptides forming enzymatic complexes in *S. cerevisiae* was obtained by combining available information in YeastCyc (65) and Complex Portal (150). Turnover numbers (k_{cat}) were retrieved from BRENDA database using functions provided by GECKO (38).

3.4.3 Allocation data and constraints

We created the ETFL models with either a constant or growth-dependent biomass composition. In the case of constant biomass composition (cE(T)FL), we used the macromolecular fractions from the biomass reaction of the FBA model (Yeast8). The mass fractions for different macromolecules were calculated using the following expression:

$$f_k = \sum_{i \in M_k} \eta_i \text{MW}_i. \quad (3.1)$$

For each type of macromolecule M_k , $(\eta_i)_{i \in M_k}$ is the stoichiometric coefficients of metabolites belonging to this macromolecule class in the biomass reaction, and MW_i their molecular weight. For example, to find the protein fraction f_{Prot} in the biomass, the stoichiometric coefficients of individual amino acids were multiplied by their molecular weight to find their mass fractions in the biomass. The sum of these amino acid ratios indicates how much of the biomass is protein. By definition, the weight of biomass should be 1 g (151, 152), which can be written:

$$\sum_{i \in \text{reactants}} \eta_i \text{MW}_i - \sum_{j \in \text{products}} \eta_j \text{MW}_j = 1. \quad (3.2)$$

When generating an ETFL model, it is important remove protein and RNA metabolites from the biomass equation to prevent double-counting of the metabolic requirements, since the explicit mRNA and peptide synthesis reactions already account for their respective participation in cell growth. In ETFL, we model the participation of macromolecules in the cellular biomass composition as follows:

$$\sum_{j \in \mathcal{J}} \text{MW}_j \cdot E_j = P^m, \quad (3.3)$$

$$\sum_{l \in \mathcal{L}} \text{MW}_l \cdot F_l = R^m. \quad (3.4)$$

In the above equations, where P^m and R^m are respectively the protein and RNA mass fractions, in g/g_{DW} , and E_j and F_l represent the concentration of enzyme j and RNA l in $\text{mmol} \cdot \text{g}_{\text{DW}}^{-1}$, respectively. P^m and R^m can either be constant (cE(T)FL), or variable and discretized (as in the vE(T)FL). \mathcal{J} and \mathcal{L} are the indexing sets of proteins and mRNAs in the model.

To create a vE(T)FL model, it is necessary to know the fraction of each biomass components at different growth rates. We gathered this information for the yeast by reviewing the literature (135, 153, 154) (data available on the online yETFL repository, accessible from the data availability statement). Since the data is usually reported for a few specific growth rates, we resampled it using piecewise-linear interpolation, as prescribed in the ETFL method (43).

Protein allocation Since all the cellular tasks of proteins are not considered in ME-models, ETFL defines a generic protein to represent the part of proteome that is not accounted for in the model, such as structural proteins, signaling proteins, or peptidoglycan synthesis. To realistically account for enzyme participation in the proteome, we define ϕ , the ratio of proteins that are associated to a metabolic task to total protein content of the cell. We use it to alter Eq. 3.3 and constrain further the enzyme pool of the cell:

$$\sum_{j \in \mathcal{J}} \text{MW}_j \cdot E_j = \phi \cdot P^m. \quad (3.5)$$

To find ϕ , we used the latest protein abundance dataset for *S. cerevisiae* available in PaxDB (155). In yETFL, this fraction is $0.55 \text{ g/g}_{\text{proteins}}$.

DNA The growth-dependence of the DNA abundance in the cell was modeled as proposed in the original ETFL formulation (43).

Carbohydrate, Lipid and Ions To consider the growth-dependence of the carbohydrates and lipids, we introduced the polymerization of lipids and carbohydrates in the ETFL formulation. To this end, we first defined a metabolite pool for each of these macromolecules. In Yeast8, each biomass component is attached to a pooling reaction that transforms the sum of specific metabolites (e.g. all carbohydrate metabolites) into a single metabolite pool (e.g. carbohydrate). The mass balance equation for these modeling metabolite pools is the following:

$$\forall i \in \{\text{Carbohydrate, Lipid, Ion}\}, \quad \frac{d[X_i]}{dt} = \eta_i^{\text{biomass}} \mu - \eta_i^{\text{pool}} v_{\text{pool}}. \quad (3.6)$$

Chapter 3. Emergence of the Crabtree effect in a model of metabolism and expression for *S. cerevisiae*

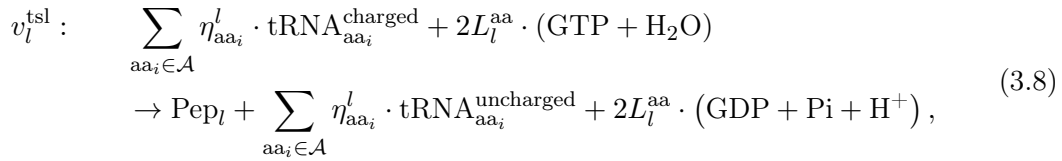
v_{pool} is the flux of through the pooling reaction, and η_i^{pool} and η_i^{biomass} represent stoichiometric coefficient of the modeling metabolite i in the pooling and biomass reactions, respectively. When it is desired to model a growth-dependent stoichiometric coefficient in the biomass reaction, the said stoichiometric coefficient can be redefined as a function of μ and calculated as follows:

$$\forall i \in \{\text{Carbohydrate, Lipid, Ion}\}, \quad \eta_i = \eta_i^{\text{ref}} \frac{\sum_{u \in \mathcal{U}} \lambda_u \cdot X_{u,i}^m}{X_{\text{ref},i}^m} \quad (3.7)$$

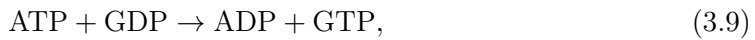
In this equation, η_i^{ref} is the stoichiometric coefficient of the pool i in the biomass reaction. $X_{\text{ref},i}^m$ and $X_{u,i}^m$ are the mass fraction of component i respectively in the original biomass equation and its discretized mass fraction at the discretized growth state number u , following notations from Salvy *et al.* (43). λ_u are binary variables activated one at a time depending on the growth rate, and \mathcal{U} is their indexing set. Because of λ_u and the formulation of ETFL, only one term of the sum will be non-zero at a time. The growth-dependent fraction $X_{u,i}^m / X_{\text{ref},i}^m$ behaves as a rescaling parameter of the stoichiometric coefficient, and allows to have a growth-dependant contribution to the biomass reaction.

3.4.4 Modifying the growth-associated maintenance

The energetic cost of growth, including maintenance of the cell and polymerization of the macromolecules, is quantified in genome-scale models with the growth-associated maintenance (GAM) (156). In ETFL, we consider the energetic cost of protein synthesis explicitly, and this cost should be removed from the GAM to avoid the overestimation of energetic requirements in the polymerization of peptides. The peptide synthesis is modeled as follows (Eq. 3.8):



where aa_i is the i^{th} amino acid, $\eta_{\text{aa}_i}^l$ represents its count in the l^{th} peptide Pep_l and L_l^{aa} is the length of the peptide in amino acid. Since 2 mol of GTP are needed to attach 1 mol of amino acid to the peptide (Eq. 3.8), and from the following interconversion reaction:



1 mol of ATP is required to produce 1 mol of GTP, we can deduce that peptide polymerization requires 2 mol of ATP per 1 mol of amino acid. We also know that the stoichiometric coefficients of amino acids in the biomass reaction of Yeast8 give information on how much of each amino acid is required to produce 1 g of biomass. From there, we compute the total amount of amino acids ($4.1 \text{ mmol} \cdot \text{g}_{\text{DW}}^{-1}$) required for the production of 1 g of biomass. Altogether, the energetic cost of peptide synthesis amounts to $2 \times 4.1 = 8.2 \text{ mmol}_{\text{ATP}}/\text{g}_{\text{DW}}$, which we remove from the GAM.

3.4.5 Gene-protein-reaction coupling

Coupling the reactions in metabolic networks with their enzymes is the most important step in the process of creating an ETFL model. Ideally, assigning enzymes to reactions requires information about: (i) the dependency of reaction on organism's genes which is gathered as gene-protein-reaction rules; (ii) catalytic rate constants (k_{cat}); and (iii) type and stoichiometry of the peptides assembly into enzymes. Whenever we did not have access to all required information, we used the following assumptions (Fig. 3.4):

- We assumed similar composition for isoenzymes, if composition information for only one of them was available. For example, if one of the isoenzymes is a dimer, the other is also assumed to be a dimer.
- We assumed that monomeric enzymes catalyze reactions (i) that depend on a single gene, and (ii) for which information about their enzyme composition was not available.
- If an enzyme peptide composition is identified, either from databases or by approximation, but its catalytic rate constant was not found, we set it equal to 70.9 s^{-1} , which is the median value for catalytic rate constants in *S. cerevisiae* (38).
- While the reactions that transport a metabolite from one compartment to another are associated with genes, their kinetic information is scarce. We set the catalytic rate constant of the proteins that catalyze these reactions to an arbitrarily large number. This ensures the gene-protein-reaction relationship is preserved, which is important for gene essentiality analysis.

3.4.6 Gene essentiality analysis

We used gene essentiality analysis to assess the quality of yETFL. The ETFL formulation enables single gene knockouts by blocking the flux through transcription reaction for each gene. If the model does not meet a minimal growth requirement knocking out a gene, the gene is considered to be essential. The predicted essential genes were compared against ex-

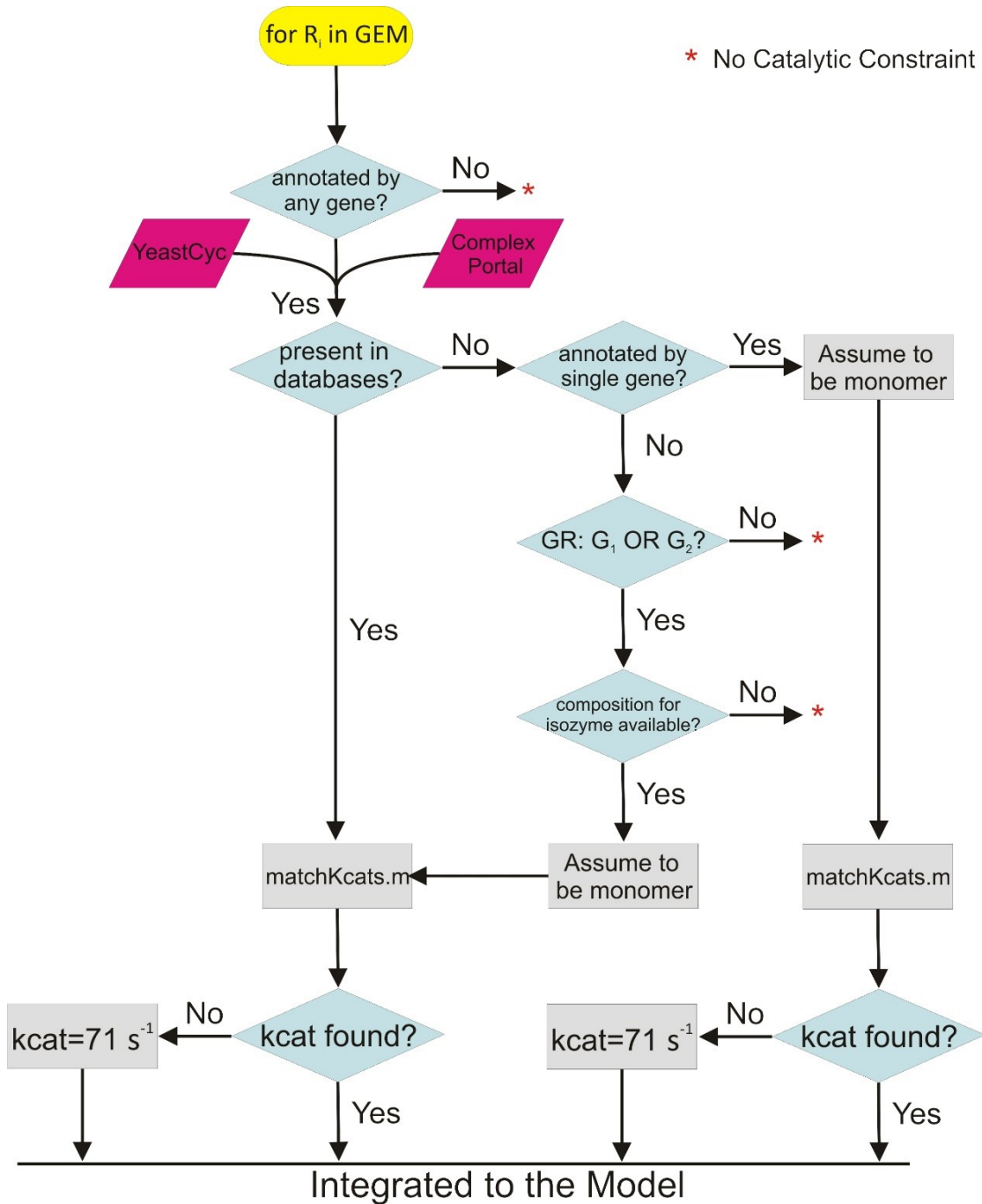


Fig. 3.4. Workflow for the integration of enzyme data into the model. The enzyme composition for the complex enzymes was found in YeastCyc (65) and ComplexPortal (150). The function `matchKcats.m` from GECKO (38) was used to find catalytic rate constants.

perimental data for *S. cerevisiae* obtained from the Stanford Yeast Deletion datasets(http://www-sequence.stanford.edu/group/yeast_deletion_project/downloads.html).

The culture medium was modified to a synthetic complete medium, as done by Lu *et al.* (133) in their essentiality studies. The Matthew’s Correlation Coefficient (MCC) was used as a metric to evaluate the quality of predictions for FBA and ETFL, because of its robustness to the imbalance in the number of essential and non-essential genes. MCC can take values from -1 to 1, where values of MCC close to -1 indicate predictions opposed to the ground truth, 0 random predictions, and 1 perfect predictions.

3.4.7 Chemostat simulations

The results of this paper were obtained by simulating the cell growth as a function of different carbon uptake rates. This allows to exhibit proteome-limited behavior and overflow metabolism in the presence of excess glucose. For all simulations (save the essentiality study, see above), the model was allowed to uptake glucose as a carbon source, some essential inorganic compounds, and molecular oxygen, as described previously in Sánchez *et al.* (38).

To capture the Crabtree effect, for different values of the growth rate ranging from 0.025 h⁻¹ to 0.41 h⁻¹, we also followed the method used by Sánchez *et al.* (38):

1. Minimization of the absolute glucose uptake rate.
2. At fixed minimal glucose uptake rate, we performed parsimonious FBA, which minimizes the total fluxes thorough the model (157).
3. We maximized the concentration of the modeling enzyme used to represent all the enzymatic activity that is unaccounted for in our model.

Additionally, we performed Chebyshev centering on enzyme variables, according to the method from Salvy *et al.* (98), to obtain a representative sample of the optimal space.

3.4.8 Code and Dependencies


The code was implemented in Python 3.7 and the commercial solver Gurobi was used to solve the MILP problems. The code relies on the ETFL (43) and pyTFA (94) packages, which use COBRApy (95) and Optlang (96). After publication, the code will be freely available under the APACHE 2.0 license at <https://github.com/EPFL-LCSB/etfl> and <https://gitlab.com/EPFL-LCSB/etfl>.

Acknowledgements

The authors would like to thank Dr. Daniel Weilandt for his valuable advice on the molecular biology of yeast, as well as important insights on the Yeast8 model. This work has received funding from the European Union's Horizon 2020 Research and Innovation Programme under the Marie Skłodowska-Curie grant agreement No 722287, and the European Union's Horizon 2020 Research and Innovation Programme under grant agreements No 686070 and 814408.

Competing Interests

The authors declare no competing interests.



4. Dose-dependent drug effect and resistance mechanisms in a cancer model of metabolism and expression

Chapter 4. Dose-dependent drug effect and resistance mechanisms in a cancer model of metabolism and expression

Pierre Salvy¹, María Masid Barcón¹, Vassily Hatzimanikatis^{1,*},

¹ Laboratory of Computational Systems Biotechnology, École Polytechnique Fédérale de Lausanne (EPFL), Lausanne, Switzerland

* Corresponding author: vassily.hatzimanikatis@epfl.ch

In this chapter, I move away from industrial biotechnology to consider the applications of computational biology in health and medicine. In particular, I consider the uses of models of metabolism and gene expression to model cell-drug interactions in the case of colon cancer. Building ME-models from human GEMs has been a long-sought goal in computational biology, but it was only thanks to the recent publication of a systematic reduction method by María Masid Barcón *et al.* (158) that I managed to generate a working model. In this chapter, I build a ME-model from a tissue-specific, systematically reduced human GEM generated using state-of-the-art methods (159, 160). I also develop a formulation to account for signaling and regulation mechanisms in the cell, using ETFL. I then use the formulation to create a model of interaction between a specific drug used in the treatment of several cancers and a colon cancer cell. I show the resulting model is able reproduce experimental results of decreased cell growth, in presence of the drug, and show the corresponding metabolic and proteomic changes. The model also highlights possible mechanisms of drug resistance the cell can use, and two growth-limiting actions of metformin occurring at different doses. Finally, I propose ways to improve such models to move towards personalized, context-aware models for personalized medicine.

This chapter is adapted from a manuscript in preparation. I designed the study, wrote the code, and performed the studies. María Masid designed the reduced model, and contributed significantly to its integration in the ETFL framework. María and I analyzed the results. María Masid Barcón, Vassily Hatzimanikatis, and I wrote the manuscript.

Once a preprint is published, all the code and documentation will be available under the APACHE 2 license at:

<https://github.com/EPFL-LCSB/tech>

<https://gitlab.com/EPFL-LCSB/tech>

4.1 Introduction

An important challenge in the treatment of cancer is the emergence of drug resistance. The plasticity of cancer cells and their strong mutation rate allows them to sometime adapt their physiology and develop a tolerance to pharmacotherapeutic agents (161). Left unchecked, this tolerance can cause the cancer to overcome its treatment(162).

Understanding how cells react to pharmaceutical stimuli is a key endeavor in drug design. Moreover, understanding how they build up resistance against previously effective therapies holds important insights on the design of more robust treatments. One explanatory variable of the cells reaction, or lack of, to a given chemical is their capacity to adapt on multiple biological levels, such as metabolism, gene expression, and global regulation (161, 162). Yet, many dimensions of these cellular states are not readily observable, let alone at a scalable level.

The necessity of tools to access the internal state of cells motivates the use of adequate modeling methods (163). In particular, models of metabolism and gene expression (ME-models) are able to capture the interactions between small molecules, metabolism, and gene expression. A recent implementation of ME-models, ETFL (43), is also suited to the integration of a several types of data, such as metabolomics, transcriptomics, and proteomics. As such, ME-models appear to be an adapted tool to study the physiology of cancer cells, and can flexibly integrate their variability into personalized, context-aware models.

Cancer mechanisms include small-molecule-protein interactions, and the deregulation of metabolism, proteome, and gene expression (164). An important component of these deregulations is the multitude of signaling cascades that change cellular behavior depending on the environmental conditions. In particular, several drugs used in the treatment of cancer act on these signaling networks to short-circuit the proliferation of malignant cells. One such drug is metformin, an ubiquitous drug for type 2 diabetes used to control blood sugar levels. Metformin was recently repurposed in the treatment of several types of cancers, occurring in liver, colon, pancreas, and bladder (165, 166).

Signaling is a keystone mechanism of cell biology, and yet state-of-the-art ME-models do not account for it. We propose in this work a formulation for regulatory interactions, using the ETFL framework for ME-models. We also model small-molecule-protein interactions, and post-translational modifications such as protein phosphorylation. We develop an implementation of a part of the regulatory network targeted by metformin and reproduce its growth-limiting action on a reduced ME-model of colon cancer cell, and show the occurrence of two modes of action at different drug doses. Finally, we highlight the changes in cell metabolism, proteome and gene expression directly and indirectly caused by metformin, and show these changes can be used to infer possible mechanisms of drug resistance.

Chapter 4. Dose-dependent drug effect and resistance mechanisms in a cancer model of metabolism and expression

The mechanism of action of metformin is multiple (167), and Fig. 4.1 reproduces here a subset of its effects, also summarized below:

1. Non-competitive inhibition of the mitochondrial NADH:ubiquinone oxidoreductase, Type I NADH dehydrogenase (respiratory complex I), curtailing ATP production;
2. Activation of AMP-mediated protein kinase (AMPK) by the changes in AMP:ATP and ADP: ATP ratios in the cell;
3. Phosphorylation and inhibition of Acetyl-CoA Carboxylase (ACCOAC) by AMPK – Lipid synthesis inhibited;
4. Activation of protein kinase A (PKA) through AMPK-mediated changes in cAMP levels;
5. Phosphorylation and inhibition of pyruvate kinase (PYK) by PKA;
6. Concurrent repression of gene expression for phosphoenolpyruvate carboxykinase (*PCK1*) by AMPK, and expression activation by PKA, through the repression of a repressor;

This is a simplified view of the multiple mechanisms with which metformin affects the cell. Rena *et al.* (167) provide a more detailed account of the physiological effects of metformin on the cell. We will use this simplified model, along with suitable modeling assumptions, to integrate in a human colon cancer cell model the regulatory and metabolic effects of metformin.

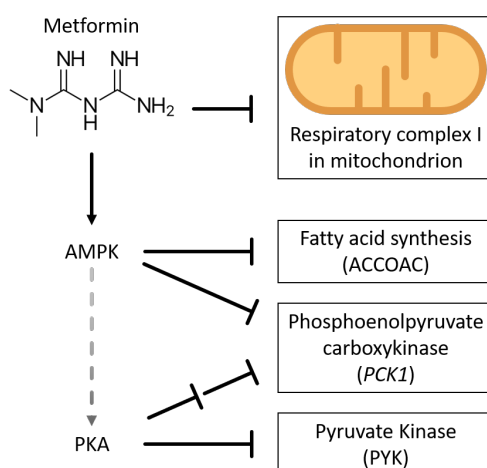


Fig. 4.1. Simplified mechanism of action of metformin on a human cell. AMPK: AMP-activated protein kinase K.; PKA: Protein Kinase A; ACCOAC: Acetyl-CoA carboxylase; *PCK1*: gene of the cytoplasmic phosphoenolpyruvate carboxykinase 1; PYK: Pyruvate kinase

4.2 Results and Discussion

4.2.1 Experimental setup

We generated a context-specific model of a colon cancer cell by integrating exometabolomics and exofluxomics data from Jain *et al.* (168) into a reduced human model (159, 160), derived from the human model RECON3D (169). We subsequently generated a ME-model from this metabolic model following the method detailed by Salvy *et al.* (43) (Table 4.1).

To integrate regulation mechanisms in the model, we considered the following assumptions:

- Metformin inhibits the respiratory complex I in a non-competitive manner (167);
- Metformin activates the expression of AMPK with Hill-type kinetics – however, *in vivo*, changes in ATP concentration activate AMPK;
- AMPK phosphorylates ACCOAC with standard Michaelis-Menten kinetics;
- AMPK activates PKA with Hill-type kinetics – however, *in vivo*, PKA is activated by AMPK-mediated changes in cAMP concentration;
- PKA phosphorylates PYK with standard Michaelis-Menten kinetics;
- *PCK1* expression is repressed by AMPK with Hill-type kinetics;
- *PCK1* expression is activated by PKA with Hill-type kinetics – however, *in vivo*, PKA represses a repressor of *PCK1* transcription.
- The cell is not allowed to produce more enzymes than the amount of the negative control to compensate for enzyme inhibition.
- Enzyme inhibition through phosphorylation decreases the catalytic rate constant of the enzyme a hundredfold.

The constraints used to model these assumptions are detailed in the section Materials and Methods. To apply metformin to a model, we follow a special sequence of optimizations, also detailed in the section Materials and Methods. The results present intracellular fluxes, their variability, and intracellular enzyme concentrations of a representative optimal solution.

4.2.2 Experimental results

To simulate the effect of metformin on the cell, we subjected the model to different concentrations of metformin, from 0.1mM to 20mM. We also provide simulation values for

Chapter 4. Dose-dependent drug effect and resistance mechanisms in a cancer model of metabolism and expression

Table 4.1. Properties of the vEFL model obtained from the colon cancer-specific reduced RECON3D.

Growth upper bound $\bar{\mu}$	0.10 h ⁻¹
Number of bins N	64
Resolution $\frac{\bar{\mu}}{N}$	0.0016 h ⁻¹
Number of constraints	18696
Number of variables	13356
Number of species	
– Metabolites	1037
– Enzymes	294
– Peptides	417
– mRNAs	417
– tRNAs	21 × 2
– rRNA	3
Number of reactions	
– Metabolic	917
– with enzymes	692
– Transport	464
– Exchange flux	373
– Transcription	417
– Translation	417
– Complexation	291
– Phosphorylation	3
– Degradation	711

a control where no metformin is added, similarly to the *in vitro* and *in vivo* experiments by Zhang *et al.* (170). Their results reprinted in Supplementary Figure 1.

The simulation predicts a decrease in cell specific growth rate as the concentration of metformin increases (Fig. 4.2-a). The negative control initially grows at a specific growth rate of 0.034 h^{-1} , while the specific growth rate of the model in presence of 20mM of metformin is of 0.018 h^{-1} . We compare these results to those of Zhang *et al.* (170) on *in vitro* and *in vivo* bladder cancer cells, and observe qualitative agreement on the decrease of cell viability at metformin concentrations higher than 1mM (see their reprinted results in Supplementary Figure 1). It is worth to note, however, that bladder cancer and colon cancer are substantially different in their pathophysiology, and the comparison of our simulated colon cancer results to *in vitro* and *in vivo* bladder cancer results mainly possess a qualitative purpose.

The first component of the action of metformin is its inhibition of the NADH:ubiquinone oxidoreductase respiratory complex I. This reaction is part of the respiratory system of the cell, and uses NADH to transfer electrons to ubiquinone (coenzyme Q10), a necessary cofactor in cellular respiration. The simulations show the feasible flux range of this reaction decreases in magnitude as the metformin concentration increases (Fig. 4.2-b). Since the size of the enzyme pool dedicated to the respiratory complex I can only be, at maximum, that of the negative control, and the catalytic rate constant of the enzyme is decreased by non-competitive inhibition by metformin, the maximal rate that the reaction can carry decreases as a function of the metformin concentration. After 2 mM of metformin, it is not optimal anymore to produce any amount of respiratory complex I (Fig. 4.2-d). Indeed, protein synthesis requires amino acids and energy, which could be used in other parts of the metabolism, rather than in the expression of an inactivated enzyme.

The reduced activity of the respiratory complex I, in turn, impedes cofactor regeneration, and has the direct effect of lowering the maximal ATP Synthase rate (Fig. 4.2-c). At 20 mM of metformin, the maximal ATP synthesis flux is at two thirds of its initial capacity, which will impede cell growth.

The inhibition of the respiratory complex I can be compensated by the cell, using the FADH₂-dependent respiratory complex II which is also able to reduce Q10 (Fig. 4.3). In particular, we observe its activation at 1 mM (Fig. 4.4-a), when the respiratory complex I carries only 50% of its initial flux. Succinate dehydrogenase is also activated, oxidizing succinate into fumarate to transfer electrons to FADH₂, which finally reduces Q10 (Fig. 4.4-b). This activation of the pathway of the respiratory complex II is not triggered by regulation, but comes from the maximization of the growth rate. It can be interpreted as a condition to reach growth optimality, and thus shows a mechanism of adaptation the cell can use to overcome one of the actions of metformin.

Chapter 4. Dose-dependent drug effect and resistance mechanisms in a cancer model of metabolism and expression

At higher metformin concentrations ($\geq 5\text{mM}$), the compensatory effect of the respiratory complex II is reduced (Fig. 4.4-a), and the flux variability of ATP synthesis does not change significantly (Fig. 4.2-c). This indicates that cellular respiration is not a growth-limiting factor anymore.

The second component of the action of metformin is the activation of AMPK. According to the model construction, at the same time as the metformin concentration increases, AMPK synthesis follows, and we observe an increase in AMPK concentration in the cell (Fig. 4.5-a). The AMPK activation follows a sigmoid shape, which is the result of the Hill-type activation used to model it. This shows the model is able to present characteristic dose-response curves, and can capture changes of metabolic and proteomic states in presence of metformin.

An important effect of AMPK is the inactivation through phosphorylation of Acetyl-

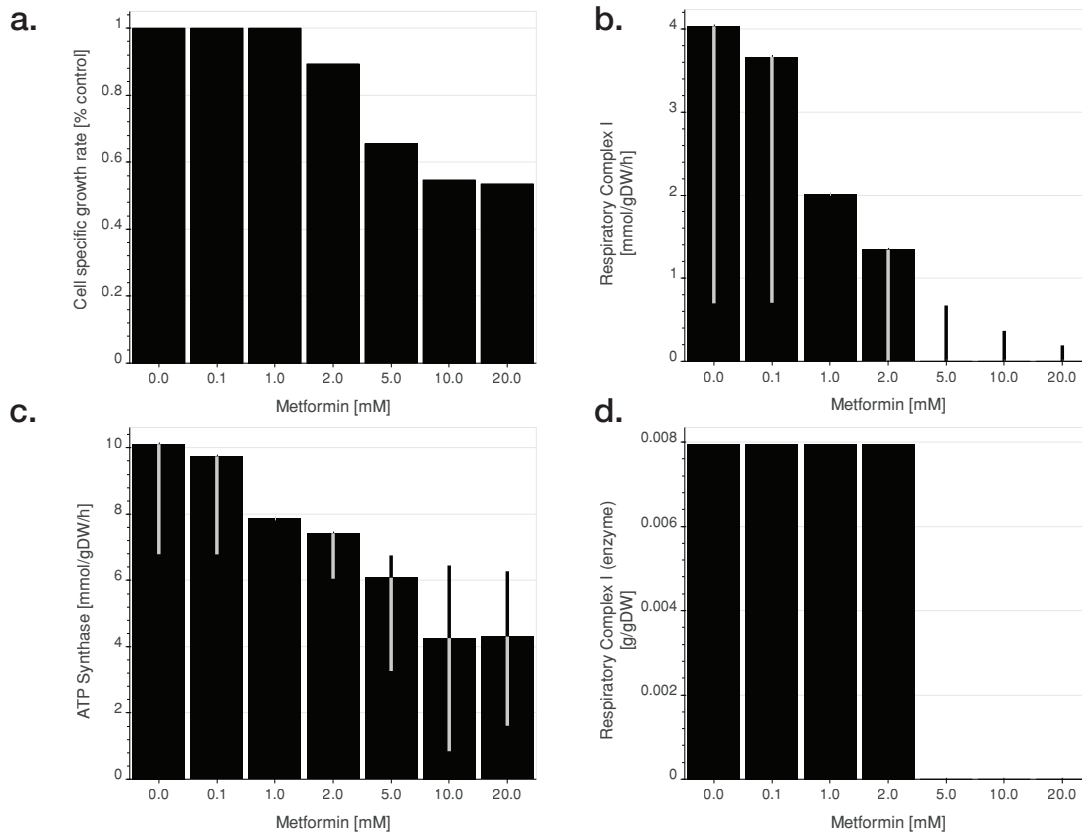


Fig. 4.2. Specific growth rate, flux rates, and enzyme levels in the colon cancer ETFL ME-model, for different metformin concentrations in mM. Fluxes have error bars that represent their variability at fixed growth rate, and the vertical bars represent their rate at the Chebyshev center. **a.** Specific growth rate of cells as a percentage of the specific growth rate of a cell without metformin treatment (negative control – left bar). The control corresponds to a specific growth rate of 0.034 h^{-1} . **b.** Flux ($\text{mmol} \cdot \text{g}_{\text{DW}}^{-1} \cdot \text{h}^{-1}$) through the NADH:ubiquinone oxidoreductase, Type I NADH dehydrogenase (respiratory complex I). **c.** Flux ($\text{mmol} \cdot \text{g}_{\text{DW}}^{-1} \cdot \text{h}^{-1}$) through the mitochondrial ATP synthase. **d.** Concentration of the mitochondrial complex I enzyme, in g/g_{DW} .

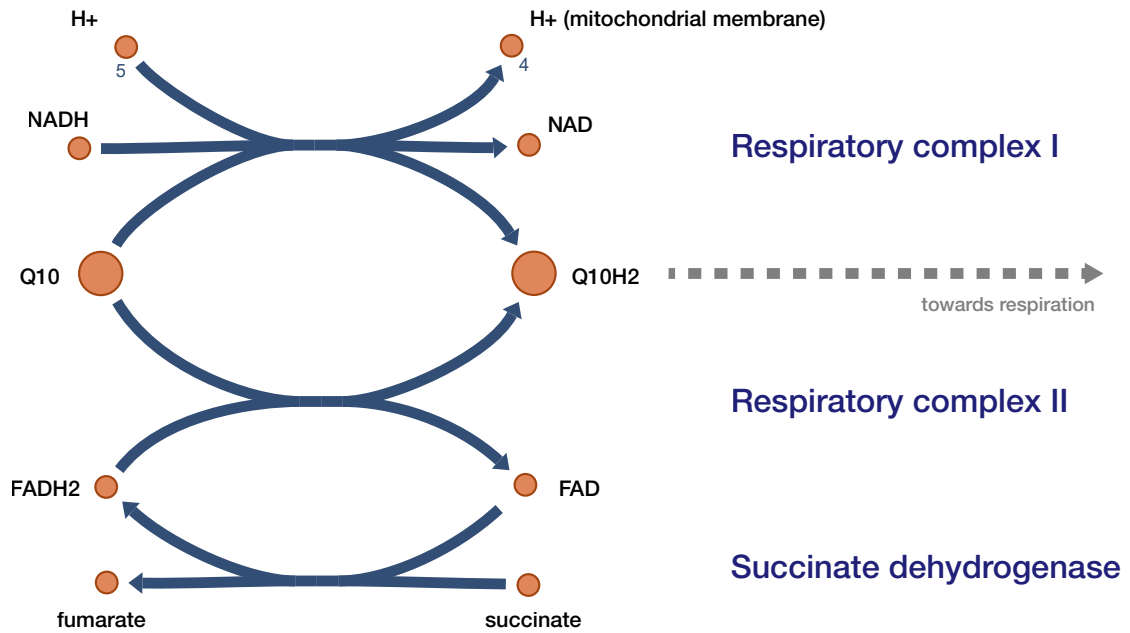


Fig. 4.3. Reactions of the respiration pathway involving the cofactors NAD/NADH, Q10/Q10H2, FAD/FADH2. The respiratory complex I uses electrons from the cellular NADH pool. The respiratory complex II uses succinate as an electron source. Q10H2 then follows the rest of the respiration pathway through complexes III and IV

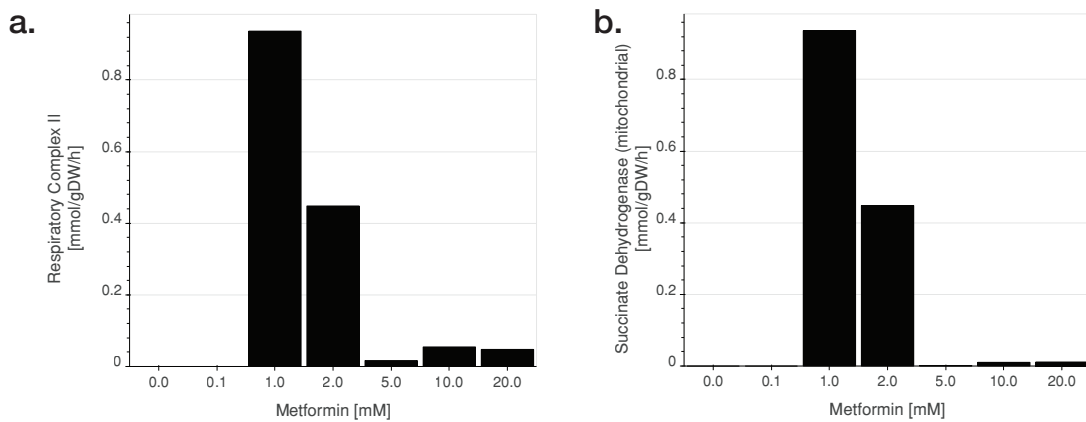


Fig. 4.4. Flux rates of reactions involved in the FADH-mediated respiratory process, at the Chebyshev center, for different metformin concentrations in mM. **a.** Respiratory complex II, transferring electrons from FADH2 to Q10. **b** Succinate dehydrogenase, which transfers electrons to FAD to form FADH2.

CoA carboxylase (ACCOAC) (Fig. 4.5-a,-c and -d). ACCOAC is responsible for the transformation of acetyl-CoA into malonyl-CoA, a necessary precursor and monomer element of fatty acids, which are responsible for the cell structure and energy storage. Under a constraint for the total amount of ACCOAC, and the progressive inactivation of its catalytic capacity, less lipids can be synthesized in the cell (Fig. 4.5-c). The reduced lipid synthesis directly impacts cell growth, since lipids are essential components of the cell, involved in cell structure and energy storage. At higher metformin concentrations

Chapter 4. Dose-dependent drug effect and resistance mechanisms in a cancer model of metabolism and expression

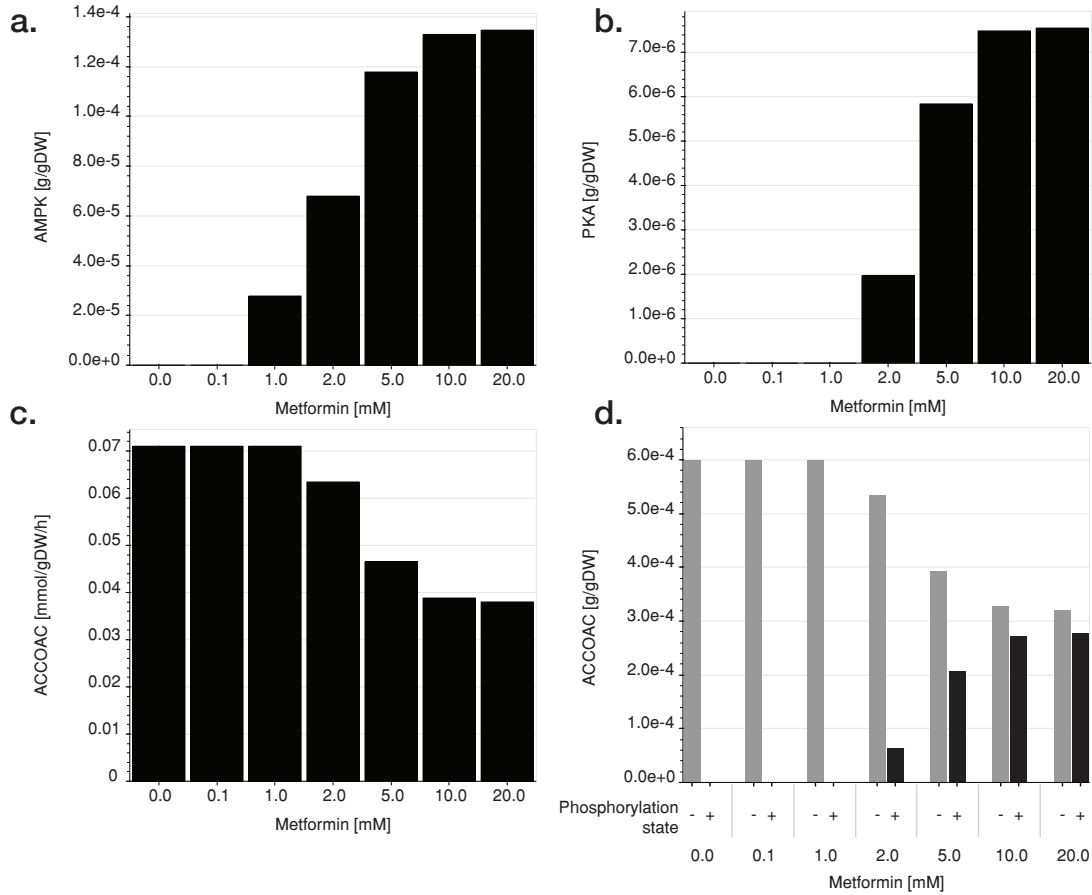


Fig. 4.5. Enzyme levels, and flux rates in the colon cancer ETFL ME-model, for different metformin concentrations in mM. **a.** AMPK concentration. **b.** PKA concentration. **c.** Acetyl-CoA carboxylase flux. **d.** Concentration of Acetyl-CoA carboxylase, (-) and (+) respectively denote the unphosphorylated and phosphorylated forms.

(≥ 5 mM), the lipid synthesis becomes a growth-limiting factor, and explains why growth continues to decrease even though the respiration of the cell is not decreasing anymore. We thus observe clearly how two different mechanisms of action of the metformin cascade contribute, at different concentrations, to the decrease of cell viability.

AMPK also activates PKA (4.5-b), continuing the cell signaling cascade. We observe that PKA also adopts a sigmoidal dose-response curve to the increased AMPK concentration, again according to the Hill-type kinetic law we used to represent it. However, we did not observe any significant effect of PKA activation on its targets, PEPCK and PYK. PEPCK and PYK do not appear to be essential to cell growth in this context-specific model, and thus their fluctuation did not impact further the cell growth.

4.3 Conclusion

We developed a ME-model for a reduced, context-aware human metabolic network, and used it to model the effects of metformin on the metabolism, proteome and gene expression of a colon cancer cell. We also developed constraints to expand the ETVL formulation of ME-models to account for signaling interactions at the level of proteins and gene expression. Our results show it is possible to include phenomena such as post-translational modifications (phosphorylation), as well as enzyme inhibition and activation, and gene expression regulation to models of metabolism and expression. With these methods, we were able to show the model reproduces experimental results of the action of metformin on tumor cells, and in particular, the growth-rate decreasing effect of metformin. The model also presents characteristic dose-response curves for the signaling proteins AMPK and PKA, and captures changes of metabolic and proteomic states in presence of a drug. In addition to recapitulating already known mechanisms, we also use the model to show the emergence of an optimal compensatory mechanism of the cell in response of the competitive inhibition of the respiratory complex I (NADH-dependent) is the rerouting of its flux to the respiratory complex II (FADH₂-dependent). This phenomenon was not modeled explicitly, but instead emerges from the model constraints, thus illustrating the predictive power of our approach.

Metformin is a molecule that belongs to the class of biguanides, known for inducing lactic acidosis (171). In parallel, the production of lactic acid via the Warburg effect is a hallmark of cancer (164), and studying the effect of metformin on this phenomenon is likely to yield important insights on toxic side effects of the treatment. Although we did not pursue it in this study, a general analysis of the production of metabolic byproducts will also be a key component of the *in silico* assessment of drug-cell interaction.

The integration of pharmacokinetics in ME-models allows the consideration of the effects of biochemical stimuli on the metabolism, regulation, and proteome of the cell. This internal cellular state is difficult to observe experimentally, and models help enrich the missing information to better understand cell physiology. Thus, it is possible to use metabolism and expression models to study drug action and eventually help in the design of new drugs. In particular, the formulation used, optimizing for growth, proposes ways in which the cell can overcome the drug-induced metabolic changes. It is amenable to testing different drug candidates, and study the optimal fitness adaptation of the cell. The model does not predict new signaling networks. However, it highlights what optimal behavior the cell might adapt under selection pressure, and this optimality is likely to be controlled by a regulation mechanism, in the same fashion as the *lac* operon controls growth optimality in mixed media in *E. coli*¹⁷ (98). These insights are valuable to design multipronged strategies to curtail the emergence of drug resistance, by anticipating potential sources of resistance before it appears.

¹⁷This is the subject of chapter 2.

One important hurdle in the future study of regulation-enabled ME-models is the size and complexity of signaling networks. We used here a limited example, with interactions only up to three steps from the signaling cascade activation. However, integrating more complete, cell-scale regulatory networks, will require additional methods. MILP-based representations of signaling networks, using binary logic, have been proposed (159). Moreover, MILP-based ME-models such as ETFL can be expanded to include such frameworks. Hence, one direct development is to integrate these two frameworks into hybrid models, with quantitative modeling of the key signaling effectors of interest, and MILP-based logic for the signaling pathways further away. Such integrated models will further improve our understanding of the complicated behavior of cancer cells, and their resistance mechanisms.

Finally, ETFL ME-models are highly suited to the integration of omics data, which are increasingly abundant in the study of cancer (168, 172). We partially pursued this feature in this study, with the integration of exofluxomics and exometabolomics. Going further, integrating transcriptomics, metabolomics, or proteomics from a tumor biopsy will directly allow to generate personalized, tissues-specific models, for several cell types. Metabolic models are already being used to elucidate cell differentiation in the immune response (142). The increased capacity of ME-models to integrate data is one step further in the direction of models tailored for cell subtypes. The versatility of ME-models will allow to tackle in a quantitative manner, for instance, cancer heterogeneity, a key element in cancer physiology and drug resistance (173). Merged with other techniques, such as community modeling, our formulation would yield a powerful tool to understand for example cellular interactions with the immune system (174), tumor-microenvironment interactions (175), or higher-order pharmacokinetic effects (162).

4.4 Materials and Methods

4.4.1 Condition-specific cancer ME-models

The colon cancer model was obtained using the reduction procedure detailed in Masid *et al.* (158, 160). The genome-scale model of metabolism used as a basis is RECON 3D (169), which was then systematically reduced around subsystems of interest: glycolysis, pentose phosphate pathway, citric acid cycle, serine, glycine, alanine and threonine metabolism, glutamate metabolism, urea cycle, oxidative phosphorylation, ROS metabolism, arginine and proline metabolism, purine metabolism, and pyrimidine metabolism. Exometabolomics and exofluxomics data from NCI-60 colon cancer cell lines (168) were then used to constrain the model in a context-aware manner.

The transformation of the model into a ME-model was done according to the standard operating procedures detailed in the original ETFL paper (43). Extra constraints were added to account for RNA Polymerase crowding, and gene copy number (see dedicated

section). In particular, we accounted for an average of 400 copies per ribosomal gene (175). Catalytic rate constants (k_{cat}) were either estimated using an average value of 172 s^{-1} (43), or taken equal to *E. coli* values when the reaction was also present in the ETFL model of iJO1366 (48, 43). Enzyme composition data was obtained from HumanCyc (65, 176). Gene-protein-reaction association rules, corrected with data from Ryu *et al.* (177), were used to estimate enzyme compositions when the information was partial.

4.4.2 ETFL

ETFL is an implementation and framework for ME-models, which allows modeling expression- and thermodynamics-enabled flux models. ETFL is a top-down formulation describing metabolites, macromolecules such as enzymes and mRNA, metabolic reactions, and expression reactions. This formulation leverages macromolecular mass balances and a liberalization scheme on top of a flux balance analysis problem (FBA) to describe intracellular metabolic fluxes, expression fluxes, macromolecular concentrations, and metabolite log-concentrations. The formulation and its equation are detailed in Salvy *et al.* (43). The core of the formulation is based on (i) mass balances written for macromolecules, including enzymes, mRNA, DNA, and tRNAs; and (ii) enzymatic catalytic constraints on reaction rates. Given a macromolecule X , its mass balances is written using a variable representing its concentration $[X]$. Under a quasi-steady state assumption, we obtain the equation:

$$\frac{d[X]}{dt} = v^{\text{syn}} - v^{\text{deg}} - v^{\text{dil}}, \quad (4.1)$$

$$= v^{\text{syn}} - k_{\text{deg}} \cdot [X] - \mu \cdot [X], \quad (4.2)$$

$$= 0. \quad (4.3)$$

where v^{syn} , v^{deg} and v^{dil} are respectively the synthesis, degradation and dilution rates of the macromolecule, μ is the growth rate, and k_{deg} is the degradation rate constant of the macromolecule.

Enzymatic reactions, found throughout the metabolism of the cell but also in its expression mechanisms (transcriptio, translation), are modeled using catalytic constraint of the form:

$$v \leq k^{\text{cat}} E, \quad (4.4)$$

where v is the flux through the said reaction, E the concentration of the catalyzing enzyme, and k_{cat} its catalytic rate constant.

4.4.3 Gene copy number and RNAP saturation

ETFL considers the saturation of a mRNA strand by ribosomes, but the original formulation did not consider the saturation of open reading frames (ORFs) of DNA by RNA polymerases.

For *E. coli*, the footprint size of the RNA Polymerase $L_{\text{RNAP}}^{\text{nt}}$ is approximately 40 nucleotides wide (BNID 107873, (83)). We assumed the human RNA Polymerase II (RNAP) has a similar footprint. To express the limit on RNAP saturation of the ORF, we can write:

$$P_l \leq \frac{L_l^{\text{nt}}}{L_{\text{RNAP}}^{\text{nt}}} G_l \quad (4.5)$$

where we recall P_l is the concentration of polymerase allocated to the transcription of the l^{th} gene, L_l^{nt} the length of the ORF in nucleotides, and G_l the concentration of ORFs of the l^{th} gene. The latter is exactly equal to the concentration of DNA times the number of copies n_l of the gene:

$$G_l = n_l \cdot \text{DNA}. \quad (4.6)$$

We can thus derive a constraint for each gene:

$$P_l - \frac{L_l^{\text{nt}}}{L_{\text{RNAP}}^{\text{nt}}} \cdot n_l \cdot \text{DNA} \leq 0. \quad (4.7)$$

4.4.4 Incorporation of signaling pathways

Three types of mechanisms must be accounted for to model the action of metformin on a colon cancer cell metabolism. In particular, metformin (i) performs non-competitive inhibition of the mitochondrial respiratory complex I; (ii) activates the synthesis of AMPK; and (iii) AMPK inactivates, through phosphorylation, ccetyl-CoA carboxylase (ACCOAC).

Upper limit on signaling compound concentrations Thomson *et al.* (178) report the concentration of different elements of the MAPK signaling cascade in *S. cerevisiae*. Concentrations range from nM to μM . To prevent the model from predicting physiologically unrealistic profiles with high concentrations of signaling peptides, we added upper bounds at 1 μM on the concentrations of AMPK and PKA.

Non-competitive inhibition Non-competitive inhibition of an enzyme was modeled using the following model:

$$v = \frac{k_{\text{cat}} [E]}{1 + [I]/K_I} \cdot \frac{[S]}{K_M + [S]}, \quad (4.8)$$

where v represents the flux of the reaction being inhibited, $[E]$ the concentration of enzyme catalyzing the reaction, $[S]$ the concentration of the substrate of the reaction, $[I]$ the concentration of inhibitor, k_{cat} the catalytic rate constant, K_M the Michaelis-Menten constant, and K_I the inhibition constant. Since $[S]$ is not always available for characterization, and in accordance with the ETFL scheme, Eq. 4.8 is transformed in the following inequality:

$$v \leq \frac{k_{\text{cat}} [E]}{1 + [I]/K_I}. \quad (4.9)$$

$$v \leq k'_{\text{cat}} [E]. \quad (4.10)$$

Finally, Eq. 4.10 shows competitive inhibition can simply be formulated as an alteration of the k_{cat} value in catalytic constraints of ETFL. This model was used for the non-competitive inhibition of the respiratory complex I by metformin.

Signaling protein activation AMP-activated Protein Kinase (AMPK) and Protein Kinase A (PKA) are two proteins that are activated during the signaling cascade provoked by the intracellular presence of metformin. To model this activation, we chose a standard Hill activation model, adapted to the ETFL formulation. In particular, the synthesis of protein peptides is modeled in ETFL with the following constraint:

$$v_l^{\text{tsl}} - \frac{k_{\text{trans}}}{L^{\text{aa}}} R_l \leq 0, \quad (4.11)$$

where k_{trans} is the maximum ribosomal translation rate constant (10 – 12aa/s for *E. coli* (69), estimated similar for humans), L^{aa} the length in aminoacids of the protein to be transcribed, and R_l is the concentration (in $\text{mmol} \cdot \text{g}_{\text{DW}}^{-1}$) of ribosomes assigned to the translation of this peptide.

To force the production of peptides, and thus of the protein of interest, we are interested in finding a lower bound to the translation flux of the peptide l :

$$f([A]) \leq v_l^{\text{tsl}}, \quad (4.12)$$

where $f([A])$ is a function monotonically increasing with $[A]$, the concentration of activator. We model this function as the product of a hill activation, and a fraction of the total ribosome translation capacity:

$$f([A]) = \frac{[A]^n}{K_A^n + [A]^n} \cdot f \cdot \frac{k_{\text{trans}}}{L^{aa}} [\text{Rib}], \quad (4.13)$$

where $[\text{Rib}]$ is the total ribosome capacity of the cell. n is the order of the Hill kinetics, K_A the Hill constant, and $0 < f \ll 1$ a non-0 small coefficient representing how much of the total ribosome capacity can be allocated to the specific translation. In this form, the expression is built with the same homogeneity as a translation flux, as detailed in ETFL. The first term of the product represents how activated the transcription is, and the product of the second and third terms represents the maximal translation flux allowed. The latter is itself a fraction of the total maximum translation flux in the cell. Since the total ribosome concentration in the cell is indirectly constrained by its growth-dependent protein requirements, we obtain a non-0 lower bound which will force a translation response when the activator molecule is present.

We can combine Eq. 4.11, 4.12, and 4.13 to define an equality constraint on the fraction of ribosomes R_l allocated to the transcription of the l^{th} peptide:

$$R_l = \frac{[A]^n}{K_A^n + [A]^n} \cdot f \cdot [\text{Rib}]. \quad (4.14)$$

Transcription activation The transcription of the phosphoenolpyruvate carboxykinase (PEPCK) is activated by PKA. We represent this interaction with scheme similar to that of the activation of signalling proteins, detailed above.

Transcription repression The transcription of PEPCK is also inhibited by AMPK. We use a standard Hill repression model to reduce the efficiency of the transcription of the PCK1 ORF by RNA polymerase.

According to the ETFL formulation, the transcription rate of the PCK1 ORF is limited

by the catalytic efficiency of the RNA polymerase and its concentration:

$$v_{\text{PCK1}}^{\text{tr}} \leq \frac{k_{\text{cat}}^{\text{RNAP}}}{L_{\text{PCK1}}^{\text{nt}}} P_{\text{PCK1}}, \quad (4.15)$$

where L_l^{nt} is the length in nucleotides of the ORF, $k_{\text{cat}}^{\text{RNAP}}$ is the catalytic rate constant of RNAP (85nt/s for *E. coli*, BNID 100060 (69), assumed to be similar in humans), and P_l the concentration of RNAP assigned to the transcription of this mRNA. We can model transcriptional repression by multiplying the catalytic rate of RNAP $k_{\text{cat}}^{\text{RNAP}}$ by a factor f . Using the classical Hill inhibition formula, we obtain:

$$f = \frac{1}{1 + \frac{Z}{K_A}^n}, \quad (4.16)$$

$$v_{\text{PCK1}}^{\text{tr}} \leq \frac{k_{\text{cat}}^{\text{RNAP}}}{L_{\text{PCK1}}^{\text{nt}}} \cdot f \cdot P_{\text{PCK1}}, \quad (4.17)$$

$$v_{\text{PCK1}}^{\text{tr}} \leq \frac{k_{\text{cat}}^{\text{RNAP}}}{L_{\text{PCK1}}^{\text{nt}}} \cdot P_{\text{PCK1}}, \quad k_{\text{cat}}^{\text{RNAP}} = k_{\text{cat}}^{\text{RNAP}} \cdot \frac{1}{1 + \frac{Z}{K_A}^n} \quad (4.18)$$

where n is the order of the Hill kinetics, K_A the Hill constant, Z is the concentration of inhibitor. $k_{\text{cat}}^{\text{RNAP}}$ is the new effective catalytic rate constant, after inhibition. Hence, transcription inhibition can simply be modeled by an alteration of the catalytic rate constant of RNAP in transcription reactions.

Enzyme inhibition by phosphorylation We model the inhibition of enzymes by their phosphorylation using three assumptions: (i) Phosphorylated enzymes have their catalytic rate constant reduced 100×; (ii) phosphorylation takes up one ATP, produces one ADP and one proton, and is catalyzed by a phosphorylating protein; (iii) the lower bound on phosphorylation follows Michaelis-Menten kinetics. Similarly to the protein activation constraint, we seek to construct a lower bound on the phosphorylation flux that will force the conversion on non-phosphorylated enzymes to phosphorylated ones:

$$f([E_A]) \leq v_{\text{phos}}, \quad (4.19)$$

$$f([E_A]) = \frac{[E]}{K_M^{\text{phos}} + [E]} \cdot k_{\text{phos}}[E_A], \quad (4.20)$$

where $[E_A]$ is the phosphorylating enzyme, $[E]$ the enzyme getting phosphorylated, K_M^{phos}

Chapter 4. Dose-dependent drug effect and resistance mechanisms in a cancer model of metabolism and expression

the Michaelis-Menten constant for this phosphorylation reaction, and k_{phos} the catalytic rate constant of the reaction. The term v_{phos} (resp. $-v_{phos}$) is added to the mass balance of the phosphorylated enzymes (resp. unphosphorylated).

If needed, it is also possible to limit the size of the phosphorylated and non-phosphorylated enzyme pool, to prevent the model to produce solutions where the small enzyme catalytic rate is compensated by a high enzyme concentration. In this case, the following constraint is added:

$$[E] + [E_P] \leq E_{tot}^{\circ}, \quad (4.21)$$

where $[E]$ represents the phosphorylated enzyme concentration, and E_{tot}° the size of the assigned enzyme pool (constant). This additional constraint was used for the ACCOAC and PYK enzymes.

Units Macromolecular concentrations in the cell are in $\text{mmol} \cdot \text{g}_{\text{DW}}^{-1}$, and kinetic parameters such as the Michaelis-Menten constant or the inhibition constant are concentrations usually reported the units mM. We convert from one to the other assuming a density of 1.08 kg L^{-1} for the cells (see Table 3, mammalian cells in Pertoft *et al.* (179)), and a drying ratio $0.5 \text{ g}_{\text{DW}} \text{ g}^{-1}$. Thus, noting X_{cell} [$\text{mmol} \cdot \text{g}_{\text{DW}}^{-1}$] and X_{aq} [mM] the concentrations respectively per gram of dried cell and per liter, the conversion is the following:

$$X_{aq} = X_{cell} \cdot 0.5 \cdot 1.08 \cdot 10^3. \quad (4.22)$$

4.4.5 Representative solution

Fluxes were described using a variability analysis (VA). At fixed growth rate, the flux through a reaction of interest is successively minimized and maximized. This allows to obtain the range of feasible flux values at the specified growth rate. It is important to look at the variability of a flux and not only at its value, because one optimal value for the objective function can possess multiple solutions.

Performing a VA on enzyme concentrations would be impractical because of the presence of a modeling enzyme, which represents the leftover fraction of proteins not directly used for metabolic functions. It essentially acts as a slack variable on the protein allocation constraint. Because of this slack, during the maximization phase of the VA, each enzyme variable will deplete the slack and take the allocated space for itself. This artifact would return a non meaningful solution, which would poorly represent the solution space.

Instead, we propose to use Chebyshev centering, as used before in Salvy *et al.* (98). The Chebyshev center is the center of the largest ball inscribed in a set of constraints. As such, it is a good candidate to represent the center of the solution space.

The Chebyshev center, can be found by optimizing a single linear problem if the solution space is a polytope (122), which is the case of ETFL problems. In the case of a polyhedron defined by inequalities of the form $a_i^\top x \leq b_i, x \in \mathbb{R}_+^n$, finding the Chebyshev center of the solution space amounts to solving the following optimization problem:

$$\begin{aligned} & \underset{r, x}{\text{maximize}} && r \\ & \text{subject to} && a_i^\top x + \|a_i\|_2 r \leq b_i \end{aligned} \tag{4.23}$$

This is similar to adding a common slack to all inequalities and maximize the its size, which maximizes the distance of the solution to the inequality constraints. However, not all variables and constraints need to be considered in the definition and inscribing of this sphere. In particular, we are interested here in a representative solution for enzyme concentrations, which only play a role in a limited set of constraints. To this effect, we define \mathcal{I}_c and \mathcal{J}_c , respectively the set of inequality constraints and variables with respect to which the Chebyshev center will be calculated. Let us also denote \mathcal{E} the set of equality constraints of the problem, a_i, c_i respectively the left-hand side of the inequality and equality constraints, and b_i, d_i their respective right-hand side. From there, we can define the modified centering problem:

$$\begin{aligned} & \underset{r, x}{\text{maximize}} && r \\ & \text{subject to} && \mu = \mu^*, \\ & && a_i^\top x + \|\mathbb{1}_{\mathcal{J}_c} \circ a_i\|_2 r \leq b_i, \quad \forall i \in \mathcal{I}_c, \\ & && a_i^\top x \leq b_i, \quad \forall i \notin \mathcal{I}_c, \\ & && c_k^\top x = d_k, \quad \forall k \in \mathcal{E}, \end{aligned} \tag{4.24}$$

where μ^* is the maximal growth rate calculated at this time step, r the radius of the Chebyshev ball, x the column vector of all the other variables of the ETFL problem, $\mathbb{1}_{\mathcal{J}_c}$ has for j^{th} element 0 if $j \in \mathcal{J}_c$, else 1, and \circ denotes the element-wise product between two vectors. Thus, $\|\mathbb{1}_{\mathcal{J}_c} \circ a_i\|_2$ is the norm of the projection of the constraint vector onto \mathcal{J}_c .

For enzymes, for example, it is akin to making the model produce more enzymes than necessary to carry the fluxes, while respecting the total proteome constraint. By maximizing the radius of the sphere inscribed in the solution space, at maximal growth rate, we are effectively choosing a representative solution of the maximal growth rate feasible space. We then use this solution as a reference point for the next computation step.

Enzymes that are strongly constrained by other parts of the model are removed from \mathcal{J}_c . In particular, AMPK, PKA, PCK2, ACCOAC (and its phosphorylated form), and PYK (2 isoforms and phosphorylated forms) are not participating in the Chebyshev center estimation.

4.4.6 Optimization procedure

The signaling cascade involves several steps where the concentrations of species will influence the synthesis of others. In order to properly account for this, it is necessary to optimize multiple times the model to successively update the concentration of all the effectors until a quasi-steady state is reached.

The model modification subroutine, given a solution, to account for the signaling pathways is the following:

- Set the metformin concentration;
- Update the lower bound on AMPK activation;
- **Optimize the model for maximal growth;**
- Fix the lower bound for the AMPK enzyme concentration;
- Update the lower bound on PKA activation;
- **Optimize the model for maximal growth;**
- Update the lower bound for the AMPK enzyme concentration;
- Fix the lower bound for the PKA enzyme concentration;
- Update ACCOAC phosphorylation by AMPK;
- Update PYK phosphorylation by PKA;
- Update PCK1 transcription regulation by AMPK and PKA;
- **Optimize the model for maximal growth;**
- Fix the growth rate;
- **Perform Chebyshev centering;**
- **Perform flux VA;**
- Release growth rate;
- Release enzyme concentration bounds on PKA and AMPK;

4.4. Materials and Methods

Table 4.2. Parameter values used in the regulation-enabled ME-model of colon cancer and their sources

Parameter	Type	Applies to	Value	Unit	Source
k_{cat}^{cplxI}	Catalytic rate constant	Respiratory complex I enzyme	275	s^{-1}	(180)
$K_{I,metf}^{cplxI}$	Inhibition constant of metformin	Respiratory complex I enzyme	1	mM	(181) & Cheng-Prusoff equation
k_{cat}^{AMPK}	Catalytic rate constant	AMPK (phosphorylating ACCOAC)	1.0	s^{-1}	(182, 183, 184)
K_A^{AMPK}	Hill activation constant	AMPK activation by Metformin	1	mM	Assumed
n_A^{AMPK}	Hill order	AMPK activation by Metformin	2	\emptyset	Assumed > 1
k_{cat}^{PKA}	Catalytic rate constant	PKA (phosphorylating PYK)	1.0	s^{-1}	Assumed similar to AMPK
K_A^{PKA}	Hill activation constant	PKA activation by AMPK	1×10^{-3}	mM	Assumed
n_A^{PKA}	Hill order	PKA activation by AMPK	2	\emptyset	Assumed > 1
k_{cat}^{PCK1}	Catalytic rate constant	PCK1, cytosol	25	s^{-1}	(185)
K_A^{PCK1}	Hill activation constant	PCK1 activation by PKA	1×10^{-3}	mM	Assumed
n_A^{PCK1}	Hill order	PCK1 activation by PKA	2	\emptyset	Assumed > 1
K_I^{PCK1}	Hill activation constant	PCK1 inhibition by AMPK	1×10^{-3}	mM	Assumed
n_I^{PCK1}	Hill order	PCK1 inhibition by AMPK	2	\emptyset	Assumed > 1
f	Maximal fraction of ribosomes	allocated to an activated enzyme	1×10^{-4}	\emptyset	Estimated with control model
$E_{tot,ACCOAC}^o$	Maximal size of the enzyme pool	(un)phosphorylated ACCOAC	5×10^{-4}	g/gdw	Estimated with control model
$E_{tot,PYK}^o$	Maximal size of the enzyme pool	(un)phosphorylated PYK	1×10^{-3}	g/gdw	Estimated with control model

- Repeat with the next metformin concentration.

Bold steps are those that involve an optimization to be run. We use this subroutine the optimization flow to update intracellular concentrations for important effectors of the signalling cascade, such as AMPK, PKA, and the activity of the respiratory complex I and reaction reactions ACCOAC, PEPCK, and PYK.

The Chebyshev solution and the results of the VA are used to find representative solutions at the end of the procedure. The figures shown in this article were all obtained after Chebyshev centering and variability analysis.

4.4.7 Parameters

A number of kinetic parameters were used to model several mechanisms of the metformin cascade. While some of them could be found using literature data, some had to be assumed at physiologically relevant values. Table 4.2 summarizes the said parameters.

4.4.8 Data and code availability

All the data used in this study and the code to reproduce it will be freely available under the APACHE 2.0 license at <https://github.com/EPFL-LCSB/tech> and <https://gitlab.com/EPFL-LCSB/tech> when a preprint of this article is online.

Supporting information

Supplementary Figure S1: In vitro cell viability after metformin treatment

Figure reprinted from Zhang *et al.* (170).

Acknowledgements

Pierre Salvy would like to thank Dr. Roeltje Maas for valuable discussions on the physiology of cancer. P.S. and V.H. were supported by the Ecole Polytechnique Fédérale de Lausanne (EPFL). M.M. was supported by the European Union's Horizon 2020 Research and Innovation Programme under the Marie Skłodowska-Curie grant agreement No 675585, and Nestle Health Sciences.

Open Science efforts

II

A method is more important than a discovery,
since the right method will lead to
new and even more important discoveries.
— Lev Landau

Une solution qui vous démolit vaut mieux que n'importe quelle incertitude.
— Boris Vian

Here is a motto I made up: Science needs ART — Accessibility, Reproducibility, and Transparency. I could also have gone with RAT or TAR, but these options seemed less marketable. In any case, I think these three items are important tenets that separate “good” from “less good”¹⁸ science. Alas, these tenets are sometimes challenged by the dull practicalities of the world.

Accessibility has been a prominent problem in academia, where many laboratories would publish results, but not the code used to obtain them. Another variant was “availability upon request”, where authors could be contacted to obtain details on the implementation of their research. Unfortunately, these interactions often ended up in rare responses from the authors, or exchanges with a characteristic response time of a year.

The reproducibility crisis has not spared computational fields, which is something I always found ironic. One would expect computer environments to be fully controlled, understood,

¹⁸I would like to avoid the word “bad”.

and reproducible from one instance to an other. Yet, any person who tried to run a piece of code found in a derelict folder from a former collaborator would agree that it is, unfortunately, seldom the case in academia. Lack of documentation, comments, but also undocumented compilers and dependencies, are a few of the many hidden parameters that should be controlled, but are often forgotten.

Finally, transparency is complicated to achieve in a competitive field where the fear of being scooped is prevalent. It is required to say what we did, but if we can avoid to say *how* we did it, the better.


Throughout my PhD, I used, developed and contributed to several pieces of software. I also wanted my software to be ART. Fortunately, the recent emphasis of grants on open-source software, open science, and efficient data management greatly helped in this endeavour. In this respect, all the code, scripts, packages I developed are available online on a variety of repository hosting services. I wrote documentation for my softwares, and actively commented my code¹⁹. All my work can be run in virtual machine containers, for which I provide build files²⁰. Finally, all of the Python dependencies I used are listed in each project directory, and most of them have version numbers attached.

This part highlights some software I have directly developed or contributed to. Some already have a significant user base outside LCSB, and some I believe, have good potential.

Software releases are often published under the form of short application notes, that highlight simple use cases and implementation in broad terms. The next chapters are under such a format.

¹⁹Comments currently make up 40% of the roughly 50 000 lines of code used to generate the results I present in this thesis.

²⁰I used Docker as a container engine and provide a **Dockerfile** for each project.



5. pyTFA and matTFA: A Python package and a Matlab toolbox for Thermodynamics-based Flux Analysis

Chapter 5. pyTFA and matTFA: A Python package and a Matlab toolbox for Thermodynamics-based Flux Analysis

Pierre Salvy¹, Georgios Fengos¹, Meriç Ataman¹, Thomas Pathier², Keng Cher Soh^{1,†}, Vassily Hatzimanikatis^{1,*},

1 Laboratory of Computational Systems Biotechnology, École Polytechnique Fédérale de Lausanne (EPFL), Lausanne, Switzerland

2 CentraleSupélec, Université Paris-Saclay, France

† Present address: Janssen Research & Development, LLC, United States

* Corresponding author: vassily.hatzimanikatis@epfl.ch

The following chapter discusses the open-source release of an important piece of software we use in the laboratory, and on which most of my research relies. pyTFA and matTFA are the first open-source, published implementations of the original thermodynamics-based flux analysis (TFA) paper (35). PyTFA is a Python package, and matTFA its MATLAB equivalent. PyTFA and matTFA add explicit formulation of Gibbs energies and metabolite concentrations to models of metabolism, which enables straightforward integration of metabolite concentration measurements. pyTFA has an active user base in academia and industry. Users sometimes reach out for explanations or contributions to the code.

This chapter is adapted from P. Salvy, G. Fengos, M. Ataman, T. Pathier, K. C. Soh, and V. Hatzimanikatis, “pytfa and mattfa: A python package and a matlab toolbox for thermodynamics-based flux analysis,” *Bioinformatics*, 2018. Georgios Fengos, Meriç Ataman, and I worked on the MATLAB code for matTFA. The three of us contributed to the documentation. I led the project in which Thomas Pathier wrote the implementation of the TFA constraints in Python for pyTFA, and I wrote all the pyTFA object-oriented system of constraints and variables. We both contributed to the documentation of pyTFA. Vassily Hatzimanikatis, Georgios Fengos and I designed the studies to perform. Georgios Fengos, Meriç Ataman and I curated the data. I made all the figures, and set up the online code repositories which contain the code documentation. I continuously updated pyTFA throughout my studies, fixing bugs, including user’s feedback, and making sure the code was easily installable through Pypi, Python’s package repository. I also set up a continuous integration system to verify the code portability, and several tutorials to reproduce the results.

All the code and documentation is available under the APACHE 2 license at:

<https://github.com/EPFL-LCSB/pytfa>

<https://gitlab.com/EPFL-LCSB/pytfa>

and:

<https://github.com/EPFL-LCSB/mattfa>

<https://gitlab.com/EPFL-LCSB/mattfa>

The content of this chapter is reproduced from the original article, with the authorization from the publisher, under the license CC-BY-NC: *“This article is available under the Creative Commons CC-BY-NC license and permits non-commercial use, distribution and reproduction in any medium, provided the original work is properly cited.”*

Abstract

Summary pyTFA and matTFA are the first published implementations of the original TFA paper. Specifically, they include explicit formulation of Gibbs energies and metabolite concentrations, which enables straightforward integration of metabolite concentration measurements.

Motivation High-throughput analytic technologies provide a wealth of omics data that can be used to perform thorough analyses for a multitude of studies in the areas of Systems Biology and Biotechnology. Nevertheless, most studies are still limited to constraint-based Flux Balance Analyses (FBA), neglecting an important physicochemical constraint: thermodynamics. Thermodynamics-based Flux Analysis (TFA) in metabolic models enables the integration of quantitative metabolomics data to study their effects on the net-flux directionality of reactions in the network. In addition, it allows us to estimate how far each reaction operates from thermodynamic equilibrium, which provides critical information for guiding metabolic engineering decisions.

Results We present a Python package (pyTFA) and a Matlab toolbox (matTFA) that implement TFA. We show an example of application on both a reduced and a genome-scale model of *E. coli*, and demonstrate TFA and data integration through TFA reduce the feasible flux space with respect to FBA.

5.1 Introduction

Constraint-based analysis on genome-scale metabolic models (GEMs) is a popular method to study metabolism and cellular physiology. Flux Balance Analysis (FBA), in particular, has been used to predict network-level behaviors, such as specific growth rate, gene essentiality, etc. The MATLAB-based COBRA toolbox (53) and its Python counterpart COBRApy (95) are today the most popular tools to perform such studies, and offer an intuitive interface to model GEMs using a linear programming formulation.

However, FBA-derived approaches often lead to flux distributions that are contradicting with physiology and bioenergetics due to the lack of thermodynamic constraints in their formulation (186, 187). We present here an implementation of Thermodynamics-based Flux Analysis (TFA) (35, 36), a framework to constrain GEMs or any metabolic network with thermodynamics. This framework allows to reduce the feasible flux solution space and eliminate thermodynamically-infeasible flux distributions, thus increasing the predictive accuracy of these models.

Previous works have been based on (35) to embed thermodynamic information in GEMs. However, they either require additional assumptions (188, 189), or calculate the ther-

thermodynamics feasibility decoupled from the FBA problem (190). TFA integrates the thermodynamics feasibility in the same MILP problem as FBA, and can unbiasedly account for all allowed thermo-dynamic profiles.

Our framework is provided under the form of a MATLAB toolbox as well as a Python package. It supported the publication of several studies integrating metabolomics in genome-scale models (191, 192, 193, 194, 195, 196)

5.2 Materials and methods

5.2.1 Embedding thermodynamic constraints

The first step towards building constraint-based models utilizing thermodynamics with TFA is to ensure a proper thermodynamic curation of the model. In particular, TFA requires the information on (1) Compartment-specific pH, ionic strength, and membrane potentials; (2) Elemental and charge balance of every reaction; (3) $\Delta_f(G'^{\circ})$ the Gibbs free energy of formation of metabolic compounds in aqueous phase, pH 7 and 0 Molar ionic strength, all concentrations held at 1M, at 25°C. (1) is obtained from literature data. If this is missing, data on phylogenetically close species can be assumed, if available. (2) is dependent on the quality of the genome scale model used as an input. TFA will however take care of adjusting the dominant protonation state of metabolites depending on their pKa and the pH of their compartment. We then perform a correction according to the Debye-Huckel equation (197) to adjust the energies to the relevant ionic strength in the compartment. For (3), $\Delta_f(G^{\circ})$ can be obtained using literature data, or estimation methods like group contribution method (87). If a metabolite does not have $\Delta_f(G^{\circ})$, the reactions that include this metabolite in their stoichiometry will not be constrained with thermodynamics. It is not possible to solely add a reaction $\Delta_r(G^{\circ})$, as the Gibbs energy needs to be linked to metabolite concentrations in order to propagate the thermodynamic constraints throughout the network. The pKa of a compound can be calculated with ChemAxon (198).

The Gibbs free energy of reactions are then transformed with respect to cellular physiology by applying the transformation as proposed in (148), as well as in this context by (36), using the given compartment-specific parameters: pH and ionic strength. Concentrations are used directly to integrate quantitative metabolomics data into the model. Upon thermodynamic curation of the model, we can formulate it as an MILP problem as explained in the supplementary information and (36).

The different types of analysis that can be performed are detailed in the Supplementary Information, and both packages include tutorials on how to perform them.

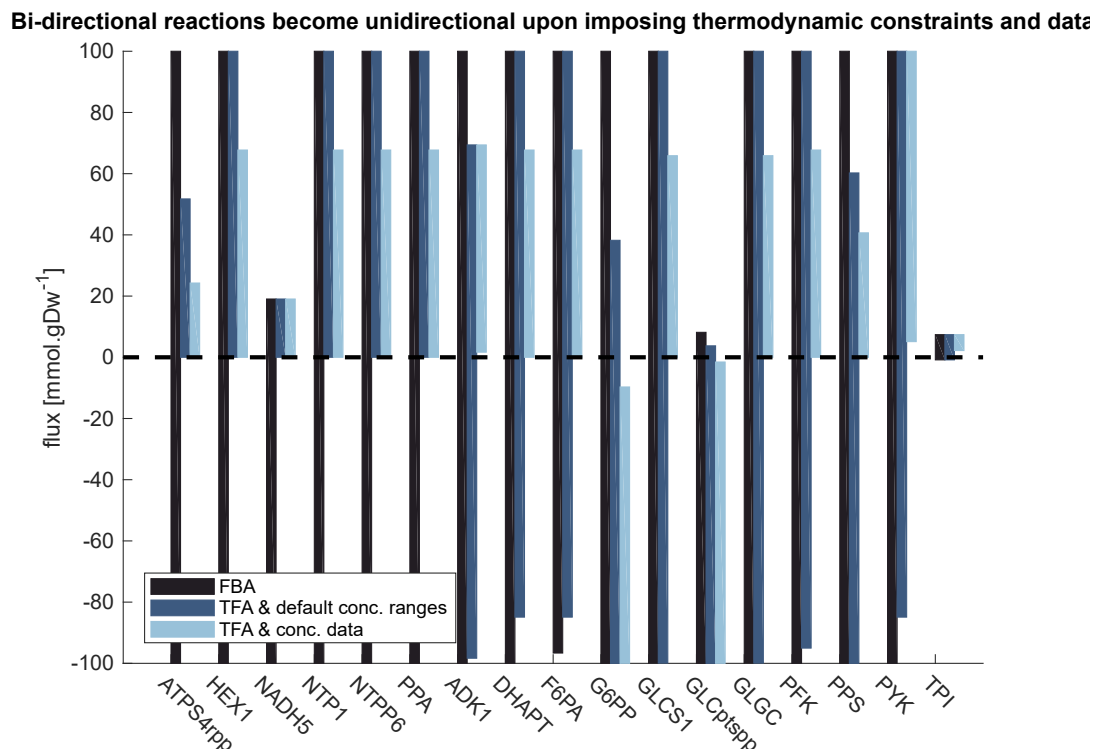


Fig. 5.1. Variability analysis for reactions whose directions are not constrained by FBA. By subsequently adding thermodynamics constraints and concentration data, all the reaction directionalities are determined.

5.2.2 Implementation

The Python package pyTFA is built to integrate with COBRApy (95), and takes advantage of Optlang (96) for solver agnosticism and model operations. The MATLAB implementation matTFA is built on top of The COBRA Toolbox (53). In the current implementation, the code uses SEED IDs (199) to match metabolites with a table of thermodynamic information taken from (87). It is also possible to input additional $\Delta_f (G^\circ)$ values manually.

5.3 Usage

The software packages come with a tutorial that demonstrates the effects of integrating thermodynamic information as well as concentration data. A reduced model of *Escherichia coli* (117), as well as the genome-scale model (iJO1366, (48)) used for its generation are provided. Figure 5.1 shows the output for a typical use case: A FBA model is constrained with thermodynamics, and then additional concentration data is added. Figure S1 illustrates that the more constrained the model is, the more reduced the allowed ranges of fluxes are. Both packages detail how to reproduce this figure


5.4 Conclusion

We propose the software package to add thermodynamic information to constraint-based metabolic models. The resulting formulation is amenable to different types of analysis with high value for the Metabolic Engineering and Systems Biology communities. We demonstrated it with a case study of a reduced system for *E. coli* focusing on glycolysis, as well as the original GEM. Our package is available for MATLAB, and Python 3, on GitHub: respectively <https://github.com/EPFL-LCSB/matTFA> and <https://github.com/EPFL-LCSB/pytfa>.

Funding

This work has received funding from the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement No 722287, RobustYeast within ERA net project via SystemsX.ch, and the European Union's Horizon 2020 re-search and innovation programme under grant agreement No 686070.

Conflicts of interest None declared



6. A Python implementation of the metabolic network analysis and reduction algorithms redGEM and lumpGEM

Chapter 6. A Python implementation of the metabolic network analysis and reduction algorithms redGEM and lumpGEM

Pierre Salvy¹, Meriç Ataman^{1,§}, Georgios Fengos¹, Benjamin Mouscadet², Romain Poirot², Vassily Hatzimanikatis^{1,*},

1 Laboratory of Computational Systems Biotechnology, École Polytechnique Fédérale de Lausanne (EPFL), Lausanne, Switzerland

2 CentraleSupélec, Université Paris-Saclay, France

§ Present address: University of Basel | UNIBAS · Biozentrum - Center for Molecular Life Sciences, Switzerland

* Corresponding author: vassily.hatzimanikatis@epfl.ch

The following chapter discusses the open-source release of another key software we use in the laboratory. redGEM (117) and lumpGEM (118) are methods for systematically reducing the complexity of genome-scale models. Reduced models are necessary in certain types of analysis where the computational complexity increases significantly with model size. This includes for example kinetic studies, where several parameters of ordinary differential equations must be estimated for each reaction in the model. Reduced models are also useful to summarize metabolism around specific physiologies, and are used in Chapter 4.

This chapter is adapted from a manuscript in preparation, P. Salvy, M. Ataman, G. Fengos, B. Mouscadet, R. Poirot, and V. Hatzimanikatis, “A python implementation of the metabolic network analysis and reduction algorithms redgem and lumpgem,” *bioRxiv*, 2020. I led the project in which Romain Poirot wrote the implementation of the redGEM algorithm, and Benjamin Mouscadet wrote the implementation of the lumpGEM algorithm. Upon completion of the project, I continued writing the integration of the code to pyTFA. The three of us contributed to the documentation. Meriç Ataman and Georgios Fengos helped with valuable contribution in the discussion in the formalization of the method and debugging of the code. Vassily Hatzimanikatis, Georgios Fengos, Meriç Ataman and I designed the studies to perform. Georgios Fengos, Meriç Ataman and I curated the data used in the example. I made all the figures, and managed the online repositories which contain the code and its documentation. I wrote most of the manuscript, and Vassily Hatzimanikatis, Georgios Fengos, and Meriç Ataman contributed to its editing.

All the related code and documentation is included in the pyTFA repository (94), under the APACHE 2 license at:

<https://github.com/EPFL-LCSB/pytfa>

<https://gitlab.com/EPFL-LCSB/pytfa>

Abstract

Genome-scale metabolic models (GEMs) provide a wealth of information when it comes to modeling microorganisms. They are key to effective and efficient metabolic engineering, and are now the basis of multiple state-of-the-art workflows (104, 33). However, their exhaustivity might come at the cost of increased computational complexity, underdetermination, sloppy parameters and poorly constrained systems, which is a problem for integrating data in studies such as for Metabolic Flux Analysis or kinetic modeling (201). As such, it is common to employ reduced models in this kind of analyses. However, reduced models often suffer from confirmation bias or selectivity bias, and might not account for important properties captured by GEMs.

The redGEM/lumpGEM framework proposes systematic analysis and reduction algorithms, using graph search and mixed-integer linear programming, to reveal the metabolic capabilities of organisms as well as to minimize bias and maximize robustness and information retention of reduced networks (117, 118). We propose here an implementation of this framework, integrated with thermodynamics-based flux analysis, which is compatible with state-of-the-art metabolic modeling software

6.1 Introduction

Genome-scale models are a cornerstone of modern computational biology. Their capacity to capture the metabolic interactions as well as gene-protein-reaction relationships makes them a powerful tool to investigate and engineer microorganisms. However, with advances in genome sequencing technology and model reconstruction, these models tend to grow ever-bigger as our understanding of the organisms they represent increases. For studies like flux balance analysis (FBA) or thermodynamics-based analysis (TFA), model size is not so much of an issue. But some types of analyses become quickly intractable as their complexity grows exponentially with respect to the network size. In particular, for kinetic modeling frameworks like ORACLE (202), which relies on parameter sampling and matrix inversions, the generation of kinetic models based on GEMs is computationally intensive for large GEMs such as the human cell model Recon3D (169), which features more than 5 000 metabolites, 10 000 reactions, each of them harboring several parameters.

To alleviate the computational burden in such analyses, reduced models are often employed (203, 204, 205, 27). However, their generation is usually done by either (i) ad hoc reconstruction of the pathway of interest, or (ii) straight subnetwork extraction from a GEM. These approaches are subject to systematic bias — Why choose this particular reaction, and not the other one? — and fail to account for unintuitive aspects of the model such as cofactor balance and alternative metabolic routes. These caveats are critical, as flux balance will behave in drastically different ways if balancing mechanisms and alternative pathways are omitted.

Chapter 6. A Python implementation of the metabolic network analysis and reduction algorithms redGEM and lumpGEM

The redGEM and lumpGEM (117, 118) algorithms were developed to respond to these issues, and allow to systematically and robustly reduce a GEM while minimizing the loss of information of the model, using a combined graph-theoretical and optimization approach.

In this work, we present our implementation of the redGEM and lumpGEM algorithms. The redGEM approach is not strictly focused only on the reduction of the stoichiometry for the generation of highly condensed network, but aims also to preserve the constitutive characteristics of metabolic networks. In particular, it retains synthesis routes for biomass building blocks in the reduced network.

Our method is provided under the form of a Python sub-module of pyTFA (94), which is a framework for thermodynamics-based flux analysis (TFA) (35, 36) and is compatible with COBRApy (95), a widely used tool for constraint-based analysis of metabolic models. The redGEM/lumpGEM framework was used to generate the backbone of kinetic models used in several integrated metabolic engineering studies (206, 195, 207).

6.2 Materials and methods

6.2.1 Reduction and lumping

The lumpGEM/redGEM framework (117, 118) reduces a GEM in two steps: (i) it calculates the n^{th} degree of connectivity between groups of reactions (subsystems) of interest, and (ii) collapses the remaining part of the metabolism in lumped reactions that fulfill the metabolic requirements of the cell.

We use redGEM's (117) breadth-first search (BFS) to calculate the n^{th} degree connections that connect the studied subsystems, yielding the core reaction network. This core reaction network will remain unchanged with respect to the original model, ensuring that all local information embedded in it is conserved.

The second part of the framework uses lumpGEM (118) as a subroutine to analyze and connect the resulting core network to the biomass reaction. Biomass reactants might be far from the core reaction network. Hence, it is necessary to compute lumped reactions that represent the subnetworks used for the synthesis of these biomass building blocks, and connect the core reaction network to the biomass reaction. LumpGEM performs this by using mixed-integer linear programming, to find sets of minimal subnetworks able to synthesize each and every biomass building block.

Table 6.1. Example lumped reactions. **3PG**: 3-phospho-D-glycerate; **AcCoA**: Acetyl-CoA; **Asp-L**: Aspartate; **CoA**: Coenzyme-A; **CTP**: Cytidine triphosphate; **DHAP**: Dihydroxyacetone phosphate; **GLCN**: D-gluconate; **PE160**: Phosphatidylethanolamine (16:0); **PEP**: Phosphoenol Pyruvate; **Pi**: Phosphate; **PPi**: Pyrophosphate; **Q8**: Ubiquinone-8; **Q8H₂**: Ubiquinol-8

Compound	Lumped reaction
L-asparate	$\text{CO}_2 + \text{NADPH} + \text{NH}_4 + \text{PEP} \rightarrow \text{Asp-L} + \text{H}^+ + \text{NADP}^+ + \text{Pi}$
CTP	$\text{ATP} + \text{GLCN} + \text{H}_2\text{O} + 3.0 \text{ NH}_4 + \text{PEP} + \text{Q8} \rightarrow 9.0 \text{ ADP} + \text{CTP} + 6.0 \text{ H}^+ + 5.0 \text{ Pi} + \text{PPi} + \text{Q8H}_2$
PE160	$3\text{PG} + 16.0 \text{ AcCoA} + 16.0 \text{ ATP} + \text{DHAP} + 15.0 \text{ H}^+ + 9.0 \text{ NADH} + 20.0 \text{ NADPH} + \text{NH}_4 \rightarrow 16.0 \text{ ADP} + \text{CO}_2 + 16.0 \text{ CoA} + 9.0 \text{ NAD}^+ + 16.0 \text{ NADP}^+ + \text{PE160} + 15.0 \text{ Pi} + \text{PPi}$

6.2.2 Thermodynamics

The package includes the option to impose thermodynamics constraints throughout the workflow, ensuring that only thermodynamically feasible subnetworks are computed. To this effect, we include in the optimization problem the constraints from thermodynamics-based flux analysis (TFA, (35, 36)).

6.2.3 Implementation

The code has been added as a submodule of the Python package pyTFA (Salvy, et al., 2018), and is built to integrate with COBRApy (95). It takes advantage of Optlang (96) for solver agnosticism and model operations.

6.3 Usage

The software packages comes with a tutorial example on the well-studied E. coli genome-scale model iJO1366 (48). In the tutorial we provide, the model is reduced around five subsystems and their the first degree connections: Citric Acid Cycle, Pentose Phosphate Pathway, Glycolysis/Gluconeogenesis, Pyruvate Metabolism, Glyoxylate Metabolism, and Oxidative Phosphorylation. From 1807 metabolites spanning 2585 reactions, the reduced model generated contains 371 metabolites spanning 599 reactions. Some examples of lumped reactions obtained in the reduction are detailed in Table 6.1

6.4 Conclusion

We propose the software package to systematically analyze and reduce constraint-based metabolic models. The reduced models capture the variability of the parent models while having a much-reduced complexity. The framework allows the use of thermodynamics constraints. Because the method is based upon state-of-the art tools, it is easily integrable into existing synthetic biology workflows. The reduced models are more amenable to computationally heavy analyses, like kinetic modeling of MFA analysis. We provide a fully detailed example reduction of the *E. coli* iJO1366 (48), under aerobic conditions and glucose feed. Our package is available for Python 3, on GitHub, as a subroutine of pyTFA: <https://github.com/EPFL-LCSB/pytfa>.

Acknowledgements

The authors would like to thank Dr. María Masid Barcón and Dr. Daniel Weilandt for their important contribution in the development of this code. They provided valuable experience and insights to improve the quality of this tool. This work has received funding from the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement No 722287, RobustYeast within ERA net project via SystemsX.ch, and the European Union's Horizon 2020 research and innovation programme under grant agreement No 686070.

Conflicts of interest None declared



7. Symbolic Kinetic Models in Python: SKiMPy

Chapter 7. Symbolic Kinetic Models in Python: SKiMPy

Daniel Weilandt¹, Robin Denhardt-Eriksson¹, Pierre Salvy¹, Ljubiša Mišković¹, Vassily Hatzimanikatis^{1,*},

¹ Laboratory of Computational Systems Biotechnology, École Polytechnique Fédérale de Lausanne (EPFL), Lausanne, Switzerland

* Corresponding author: vassily.hatzimanikatis@epfl.ch

The following chapter discusses the use and open-source release of code we developed to better accommodate kinetic models in our workflows. This software, SKiMPy, is the product of long collaborative work with my colleagues Daniel Weilandt and Robin Denhardt-Eriksson, with whom we set out to write an object-oriented framework to handle kinetic models as easily as we handle GEMs in pyTFA. SKiMPy was used by several PhD students and interns in our laboratory on a variety of projects, spanning the dynamics of the Warburg effect in cancer cells to initial rate experiments and the optimization of enzyme reaction mechanisms (data not published yet).

This chapter is adapted from a manuscript in preparation. I wrote the global architecture of the package, and defined its paradigm with dynamic declaration of classes, and symbolic handling of expressions. Daniel Weilandt and I worked on the interface to transform GEMs and pyTFA models into SKiMPy models, and the modal analysis module. Daniel Weilandt derived the kinetic expressions for the mechanisms, and refined the package based on his usage during his research work, including faster compilation of functions into machine code. Robin Denhardt-Eriksson wrote the general sensitivity analysis framework and parameter resampling. Daniel Weilandt and Robin Denhardt-Eriksson both wrote the (re)sampling and MCA modules. Daniel Weilandt generated the data presented in this manuscript. Daniel Weilandt, Robin Denhardt-Eriksson and I wrote the manuscript. Expertise on ORACLE was provided by Ljubiša Mišković. Vassily Hatzimanikatis and Ljubiša Mišković edited the manuscript.

All the code and documentation is available under the APACHE 2 license at:

<https://github.com/EPFL-LCSB/skimpy>

<https://gitlab.com/EPFL-LCSB/skimpy>

7.1 Introduction

Large scale metabolic kinetic models have become a valuable tool to bridge the worlds of computational biology and living systems. Such models are useful to understand and engineer organisms, from industrial recombinant hosts to analyzing cell-pathogen interactions and pharmacokinetics. Conventionally, those models rely on detailed information on the kinetics of individual reactions happening in the cell. However, with the limited availability of kinetic data, and the ever-growing amount of biological networks at our disposal, the building of large-scale kinetic networks is confronted with critical uncertainties (201). Additionally, parameters obtained from *in vitro* experiments are sometimes not correct to describe kinetics in physiological environments, where the medium is not well mixed and suffers effect from crowding (208, 209, 210). As a result, the fraction of available kinetic data in models tends to decrease.

In this context, the development of parameter estimation techniques emerges as a necessary tool to analyze the dynamic behavior of cellular systems. Different types of approaches have been suggested to tackle this uncertainty. These fall in two main categories: (i) parameter fitting to experimental data – these include genetic algorithms (211), Bayesian inference (212), simulated annealing (213); and (ii) parameter sampling (202). In particular, several methods building upon genome-scale models to generate large-scale kinetic models have been proposed in the last years, using parameter sampling or parameter fitting. Among them, the ORACLE workflow by Mišković *et al.* (202) has been proven to be a useful tool to estimate parameters large scale kinetic models that are in alignment with a steady state flux profile and thermodynamically feasible concentration profile (191, 201, 206, 214). ORACLE marked an important achievement in the systematic generation of large-scale kinetic models in computational biology.

We propose a Python framework for object-oriented kinetic modeling that capitalizes on a symbolic formulation of kinetic laws – Symbolic Kinetic Models in Python, or SKiMPy. This symbolic formulation allows models to be agnostic to the type of analyses to be performed. Our proposed framework is compatible with both parameter fitting and sampling methods, taking advantage of the modularity of its implementation. In particular, the ORACLE method, metabolic control analysis (MCA), (total) quasi-steady state assumption integration, parameter (re)sampling, and sensitivity analysis are implemented.

7.2 Material and Methods

7.2.1 Symbolic kinetic models

The ordinary differential equations describing a biochemical reaction network can be derived directly from the mass balance of the N reactants participating in the M reactions

of the network:

$$\forall i \in \llbracket 1, N \rrbracket, \quad \frac{dX_i}{dt} = \sum_{j=1}^M n_{ij} \nu_j(\mathbf{X}, \mathbf{p}), \quad (7.1)$$

where X_i denotes the concentration of the chemical i , n_{ij} is the stoichiometric coefficient of reactant i in reaction j and $\nu_j(\mathbf{X}, \mathbf{p})$ is the reaction rate of reaction j as function of the concentration state variables $\mathbf{X} = [X_1, X_2, \dots, X_N]^\top$ and the parameters $\mathbf{p} = [p_1, p_2, \dots, p_L]^\top$. The functions $\nu_j(\mathbf{X}, \mathbf{p})$ are the given rate laws of their respective reaction.

Within this framework we distinguish between two different types of rate-laws: (i) elementary rate laws that are based on molecular interactions; and (ii) apparent rate-laws that phenomenologically describe the reaction rate. These apparent rate laws are strongly dependent on the assumptions made on the mechanism of the reaction. For enzymatic reactions it is commonly assumed that the enzyme quantity is conserved and the enzyme complex concentration is in a quasi-steady state. These assumptions allow to simplify, for each reaction, the elementary reaction rate laws to a single Michaelis-Menten rate law. An overview of the implemented mechanisms and their respective assumptions is given in Table S1 in the Supplementary Data.

7.2.2 Sampling steady state consistent parameter sets

Large-scale kinetic models often suffer from a lack of data to calibrate their parameters (201). We approach the problem by sampling unknown parameters. In particular, for the i^{th} species concentrations $[S_i]$ as well as its Michaelis-Menten constant in the j^{th} reaction K_M^{ij} , we use the transformation proposed by Mišković *et al.* in the ORACLE workflow (202):

$$\sigma_{ij} = \frac{[S_i] / K_M^{ij}}{1 + [S_i] / K_M^{ij}}. \quad (7.2)$$

This reformulation allows to replace the unbounded sampling of concentration ranges and Michaelis-Menten constants by a sampling on the $[0, 1]$ interval.

For every parameter sample the Jacobian of the dynamic system is calculated according to the formulation proposed by Wang *et al.* (215). The inverse of the real part of the eigenvalues of the Jacobian give the characteristic time constant of the envelope of the response of the linearized system. Because of this, a Jacobian whose largest eigenvalue has a positive real part will yield unstable models. Also, eigenvalues with a negative real part too close to 0 will yield slow dynamics, potentially slower than metabolism. As a result, it is necessary to subsequently filter the parameter sets to discard (i) unstable

models and (ii) models with slow dynamics. This step might be limiting in some cases. In our experience across different organisms and metabolic networks, from 30% to only 0.01% of the samples were stable models.

7.2.3 Serialization

We observed that the development of kinetic models was subject to iterative refinement, for instance defaulting all the reactions to Michaelis-Menten mechanisms before implementing Hill kinetics for known enzymes. This was often performed through the tedious editing of undocumented spreadsheets without standards. To alleviate this issue, models are serialized in the YAML format, a human-readable data serialization language which retains the object specifications of SKiMPy. Refinements to the models can be done through a simple text editor in a typed and controlled environment enforced by the YAML specifications.

7.2.4 Implementation

SKiMPy is a Python package provides an object-oriented interface to construct the symbolic expressions using the Python package `sympy`. SKiMPy also precompiles these expressions into machine code using Cython. SKiMPy further integrates the SUNDIALS ODE-Solver package (216) using the interface provided by the package ODES (217).

7.3 Usage

The software package comes with different tutorials demonstrating (i) how to build reaction network with elementary and apparent reaction rate laws; and (ii) how SKiMPy can be used in combination with an implementation of thermodynamics-based flux analysis (pyTFA, (94)) to sample parameters following the ORACLE workflow.

As an example of SKiMPy workflow to characterize a kinetic model, we provide a pyTFA model of the core metabolism of *E. coli*, derived from Varma *et al.* (27). The TFA model provides a steady-state solution that maximizes the cellular growth rate subject to stoichiometric and thermodynamic constraints. The kinetic parameters of the model were sampled using SKiMPy. In the next step, we perform an analysis of the basins of attraction; therefore, one set of parameters is chosen and integrated for 100 different initial conditions. These initial conditions are sampled to allow for concentration five fold larger and five fold lower than the reference set (Fig. 7.1-a). Fig. 7.1-b shows the euclidean distance of each perturbed model to the reference concentration. Some perturbed samples revert to the original steady state, while others switch to a new steady state. This clearly shows the existence of two steady states, and some perturbation experiment allow the model to cross from one basin of attraction to another. Fig. 7.1-c

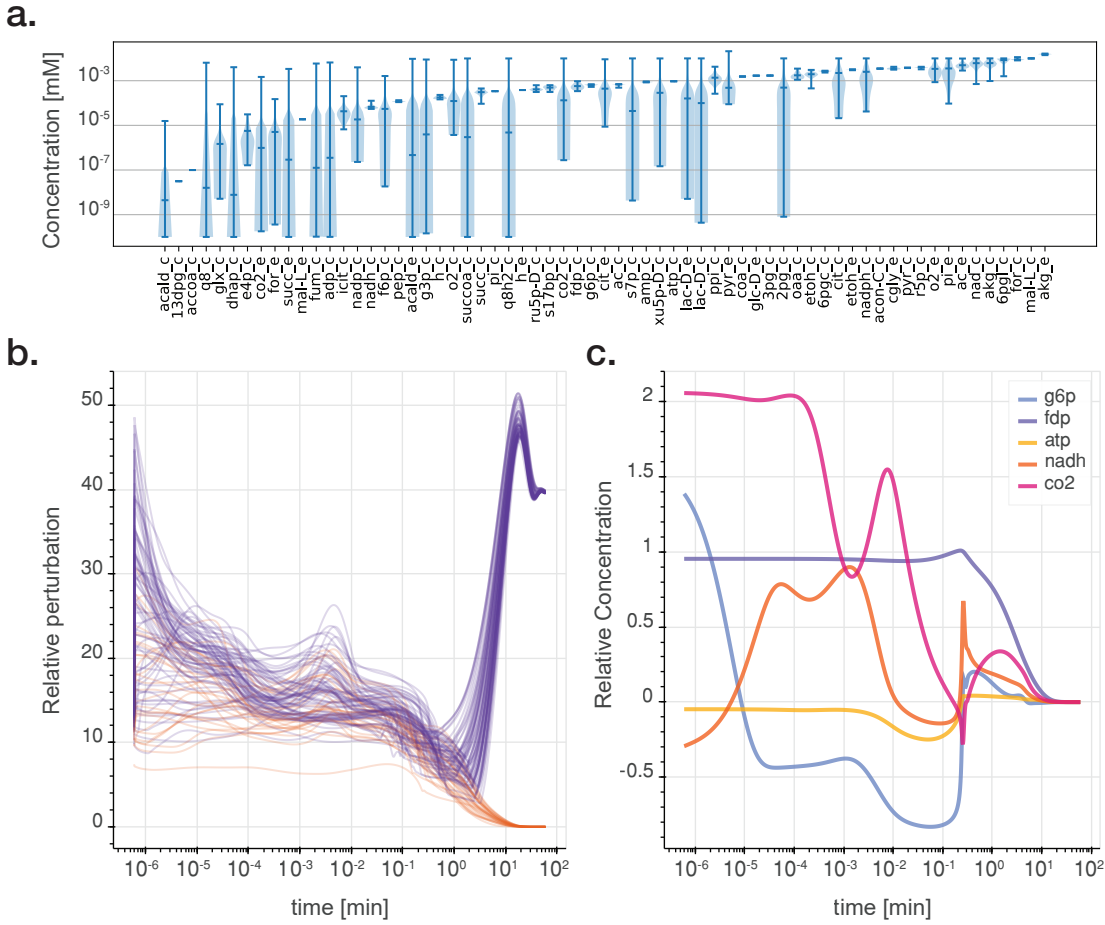


Fig. 7.1. Different outputs from Skimpy. **a.** Violin plot of the distribution of the 100 sampled concentrations for the metabolites in the model. **b.** Euclidean norm of the relative deviation of the 100 different resulting perturbation experiments. Each line represents the norm of the relative change of concentration over time versus the reference concentration. Orange lines converge towards the reference steady state, purple lines towards a new steady state. **c.** Relative concentration deviation from the reference state, in a specific perturbation experiment, with a highlight on the concentration of glucose 6-phosphate (G6P, light blue), fructose 1,6-biphosphate (FDP, dark blue), ATP (yellow), NADH (orange) and CO₂ (pink).

shows one particular perturbation experiment for which the concentrations of central metabolites are singled out.

7.4 Conclusion

We developed a framework that is able to model biochemical systems in a fashion that is agnostic to which type of analysis it will be subjected to. We show that such a model is amenable to a range of studies, without additional transformation. We demonstrate its usage on two models, one of which being a reduced genome-scale model. Such a tool will be key in the development and analysis of large-scale kinetic models.

Our package is available for Python 3 on GitHub (<https://github.com/EPFL-LCSB/skimpy>) and Gitlab (<https://gitlab.com/EPFL-LCSB/skimpy>).

Acknowledgements

The authors would like to thank Dr. Georgios Fengos for his valuable help on kinetic modeling. This work has received funding from the European Union's Horizon 2020 Research and Innovation Programme under the Marie Skłodowska-Curie grant agreement No 722287, the European Union's Horizon 2020 Research and Innovation Programme under grant agreements No 686070 and 814408, and the Ecole Polytechnique Fédérale de Lausanne (EPFL).

Concluding remarks

Conclusions

Metabolic engineering is a young field with potential applications in important domains, including health, food, and industrial biochemistry. The understanding and engineering of living cells requires new tools, be it for whether human cells, plants, or industrial hosts such as *E. coli* or *S. cerevisiae* are concerned. In this thesis, I provided a new framework to design models of metabolism and gene expression (ME-models), as well as new methods to analyze them. The methods I presented allow to (i) integrate experimental data to characterize observed cellular physiologies; (ii) elucidate non-measurable cellular states underlying observed cellular physiologies; (iii) predict, independently of experimental data, cellular responses that are empirically validated. I also provided details on several computational tools I have developed or contributed to, that can be used to deconvolute and better model biological systems.

In the first part of this dissertation we explored a new formulation of ME-models, ETFL. I derived the whole formulation from the cell biochemistry, and applied it to several organisms. I showed the formulation allows to capture complex phenotypes emerging from growth optimality, in particular under the constraint of proteome limitation. I also provided a framework for the integration of proteomics, transcriptomics, and metabolomics data in the models.

In **Chapter 1**, I tackled the problem of efficient ME-models accounting for thermodynamics. I derived a full set of equations representing cellular metabolism and gene expression, using biochemical knowledge, and provided a transparent bilinear formulation for ME-models, ETFL. I then discretized the bilinear problem in piecewise-linear MILP, and applied the formulation to a model of *E. coli*. The introduction of integers is also an opportunity to integrate thermodynamic constraints to the model, using thermodynamics-based flux analysis (TFA) (35, 36). The formulation provides a transparent way to integrate proteomics and transcriptomics through the introduction of protein and mRNA concentrations, and metabolomics through TFA. I showed the model accurately simulates protein-limited growth, allows to calculate feasible proteomes and transcriptomes, and that it captures well gene essentiality information. Finally, I also pointed out the solving

Conclusion

performance of the problem is on par if not better with the state of the art.

In **Chapter 2**, I used ETFL to model the diauxic growth of *E. coli* in a batch reactor. I first devised a conceptual model to show analytically that the sequential use of carbon sources in a mixed carbohydrate medium is the result of an optimal program for growth under the constraints of proteome allocation. Then, I set out to show the result generalizes to genome-scale models. To do so, I designed a dynamic ME-model method (dETFL) that accounts for time-dependent variation of enzymes and mRNA concentrations in the cell. I then validated dETFL by simulating the growth of *E. coli* on glucose, and showed that dETFL predicts acetate reconsumption, a diauxic behavior, in quantitative agreement with experiments. Finally, I used the dynamic ETFL formulation to simulate the growth of *E. coli* on a mix of glucose and lactose. I showed that, in accordance with the conceptual model, the dETFL model predicts the preferred consumption of glucose over that of lactose. I also showed that initial conditions change the cell fate, as cells precultured in lactose will adapt their proteome to consume glucose first, and then glucose. This allows us to postulate the regulation mechanisms of diauxie are an emerging control system to ensure the optimality of the growth of *E. coli*.

In **Chapter 3**, we applied the ETFL framework to build the first genome-scale ME-model of a eukaryotic organism, *S. cerevisiae*. We performed gene essentiality checks to validate the model accuracy, and proceeded to show the model is able to reproduce the phenotype of overflow metabolism. Indeed, under excess glucose, the model predicts fermentative processes to happen, under the form of ethanol secretion (Crabtree effect). The model also independently and quantitatively reproduces experimental data from fermentors, validating the use of ETFL to model eukaryotic organisms in the context of industrial biotechnology.

In **Chapter 4**, I built a ME-model from a context-aware reduced human model, for a cell line of colon cancer. This is the first time a ME-model formulation has been used on a human model. I supplemented this model with a method to implement a partial signaling cascade in the context of ME-models, which is an important milestone as signaling cascades are key components of cancer physiology. Using this method, I showed the model is able to quantitatively reproduce the deleterious effects of the antidiabetic drug metformin on tumor cell growth, and I showed in particular two different modes of actions that limit cell growth at different doses. I also show that the model can be used to find mechanisms of resistance that can be used by the cell to evade one therapeutic action of metformin. The combination of context-specific reduced models and ME-model formulation provides a powerful platform to create personalized, context-aware models of metabolism and gene expression and regulation. Such models can be used to design and evaluate drug treatments, and are a important milestone on the way towards personalized medicine.

The second part of the dissertation provided details on my contributions in computational

biology, under the form of software packages.

Chapter 5 detailed pyTFA and matTFA, two toolboxes (in Python and MATLAB) to perform thermodynamics-based flux analysis (TFA) on genome-scale models of metabolism. The code is provided with a documentation, tutorials, examples, and models ready to be used. TFA is a method that greatly improves the quality of the solutions given by normal FBA models (35, 36), and providing an open source implementation of the method is valuable for the community of computational biology.

Chapter 6 described an implementation of the algorithms redGEM and lumpGEM (117, 118), two important methods to systematically reduce genome-scale models around pathways of interest. These methods allow the construction of models that still capture an important fraction of the metabolism, but with a reduced number of reactions and variables. Such reduced models play a vital role for analyses where complexity matters, such as the construction of kinetic models (201, 186).

Chapter 7 finally described a framework to construct symbolic kinetic models, SKiMPy. The models can be generated from genome-scale models or other types of biological networks. Due to its symbolic formulation, SKiMPy is able to automatically transform the specified kinetic system under a form suitable for different types of analysis, such as metabolic control analysis, sensitivity analysis, or time integration. The easy handling of kinetic models for several types of study is an important step towards the systematic reconstruction of context-aware kinetic models of metabolism, an important goal for systems biology (42, 201).

Outlook

Cellular systems are complicated. In this work, I propose some new, and some improved methods to model cells at level of the metabolism and genome expression. Yet, the road is still long, and there is still much to build and discover in metabolic engineering, and systems biology in general.

In this dissertation we mostly looked at quasi-steady state models, and to some extent an approximation of a dynamic formulation using quasi-steady state assumptions. Yet, a lot of the physiology of the cell relies on transient responses, be it in cell signaling (quorum sensing, immune response), or very literal cell physiology (muscular response, neuron activation). Systems biology at steady state is extremely powerful for industrial biotechnology and fermentation processes, but it is only the doorstep to a wider, wilder field in which we only have rudimentary tools. An important milestone on the way to a more sophisticated systems biology is whole cell models. Reduced, summarized models have been successfully reproducing experimental data (81, 82, 211) , and even a whole cell model of *Mycoplasma genitalium* was published (218). However, truly integrated

Conclusion

genome-scale whole cell models are not common yet, for reasons that include the difficulty of model building, the computational time, and the difficulties in integrating experimental data (219). I believe that the ETFL formulation, with its systematic construction of the genome expression system, improved computational efficiency, and facilitated 'omics integration, is a step in the direction of accessible whole-cell models.

Models of metabolism and genetic expression have the potential to impact several important topics in systems biology. In the context of industrial biotech, ME-models and methods derived from ETFL are of particular relevance.

ME-models for platform hosts in industrial biotechnology, such as *E. coli*, *S. cerevisiae*, or *Y. lipolytica*, may provide a strong analysis framework to analyze experimental data but also produce informed engineering decisions to improve the yield, productivity, and health of the cells (220). The ability of ME-models to bridge genotype and phenotype is key to properly interpret the effects of genetic engineering, and help decision-making for the iterative engineering of these microbes.

The modeling of communities of microorganisms has garnered a lot of attention (74). Synergistic interaction between microorganisms who have co-evolved might allow to bypass many steps of otherwise painful microbial engineering, and they are the mechanisms behind a fair share of today's alimentation, including cheese (221), wine (222), or kefir (223). The integration of (reduced) ME-models to the study of these communities may be a way to better understand complex phenotypes arising from the trade-off between shared resource allocation, and global resource limitations – in other words, the price of anarchy in winemaking.

Close to the subject of microbial communities is that of metagenome analysis. Progress in metagenome sequencing now allows the systematic reconstruction of the genome-scale models of hundreds of microorganisms in a single sample of gut microbiome (44). Using the standard operating procedure I designed in ETFL, one could imagine also a systematic ME-model reconstruction for each organisms in a metagenomic sample. Such models can inform on the organisms in presence, how to cultivate them, and how to engineer ways to change their interactions to durably modify gut microbiome.

We showed with dETFL that cellular genetic regulation could be understood as a control system to ensure cell optimality with respect to its (hidden) objective function. Genetic regulation is an important part of the cellular physiology, and constraint-based models can help understand the complex, redundant and multi-layered interactions between metabolism and gene expression. This interaction of paramount importance when dealing with cancer tissue, for example, as one of the the hallmarks of the cancer cells is the global deregulation of their metabolism and signaling pathways (164, 159).

Finally, in the context of health and medicine, easy integration of omics data in ME-models will play an important role in achieving context-aware, personalized models of metabolism

and expression, which can subsequently be used for personalized medicine. Each cancer is different, and even the tumor micro-environment shows a strong heterogeneity (172, 224). This heterogeneity can be accounted for with the efficient integration of data from biopsies into our models. Other cellular systems from the human body are also of interest, such as the differentiation of T-cells in immune responses. Metabolic studies of this phenomenon already promise new insights on the mechanisms of immunity (142), and ME-models will allow even better characterization of the genomic and metabolic states of T-cells. Deep insights on the inner workings of immune cells have a strong value for research in the context of immunity-related diseases, which includes several types of dementia, genetic diseases, HIV, and cancers.

ME-models have more to say. Industrial biotechnology, food science, and health and medicine are chief fields in which they have a voice, and it is clear their democratization is important to tackle the challenges Systems Biology has to face in the 21th century.

★
★ ★

It has been said that a “wise and imaginative perception and formulation of critical questions and problems” was necessary for a successful integration of mathematical approaches in biotechnology (225). This thesis attempted to deliver some imaginative formulations, maybe even some clever tricks, to achieve higher levels of perception of what happens inside the cell, in a multitude of contexts. Wits and imagination now need to be scaled up, to tackle problems in systems biology of the next order of magnitude.

Appendix

III



A. ETFL: Supplementary information

Supplementary Information

The ETFL formulation allows multi-omics integration
in thermodynamics-compliant metabolism and
expression models

Salvy et al.

Nondimensional Scaling

A critical issue in the formulation of this problem is the different orders of magnitude the variables belong to. Fluxes are typically between $10^{-3} - 10^1 \text{ mmol.gDW}^{-1}.\text{h}^{-1}$. Protein concentrations are around $10^{-6} - 10^{-3} \text{ mmol.gDW}^{-1}$, and mRNA concentrations $10^{-10} - 10^{-6} \text{ mmol.gDW}^{-1}$. The relationship between these scales is given by the catalytic rates of enzymes and expression machinery, which span $10^3 - 10^6 \text{ h}^{-1}$. As a consequence, the constraint matrix becomes ill-conditioned and the solver has to operate close to, or sometimes beyond, their maximal solving accuracy (usually around 10^{-9} for commercial solvers such as ILOG CPLEX or Gurobi)

In order to circumvent these limitations, we operate a scaling of the EP which will reduce the numerical difficulty of the problem.

In particular, we consider nondimensionalization by upper bound as a method, which will also allow to reduce the effective range of the variables seen by the solver.

Page 2 details the full bilinear EP formulation, before and after scaling. Page 3 introduces nondimensionalization constants. Page 4 shows the relationship between original variables and scaled variables. Page 5 summarizes upper bounds used for nondimensionalization, and the nondimensionalization variables.

Appendix A. ETFL: Supplementary information

<i>Initial</i>	<i>Scaled</i>
$S \cdot v = 0$	(FBA)
<i>Catalytic constraints</i>	
$v_j^f - k_{cat,f}^j E_j \leq 0$	$\tilde{v}_j^f - \tilde{E}_j \leq 0$ (FC) _j
$v_j^b - k_{cat,b}^j E_j \leq 0$	$\tilde{v}_j^b - \tilde{E}_j \leq 0$ (BC) _j
<i>Macromolecule mass balances</i>	
$v_j^{asm} - v_j^{deg} - \mu * E_j = 0$	$\tilde{v}_j^{asm} - \tilde{v}_j^{deg} - \frac{\mu}{k_{deg}^j} * \tilde{E}_j = 0$ (EB) _j
$v_l^{tcr} - v_l^{deg} - \mu * F_l = 0$	$\frac{1}{k_{deg}^l F_l^{ub}} \cdot \frac{k_{cat}^{RNAP} P_l^{ub}}{L_l^{nt}} \cdot \tilde{v}_l^{tcr} - \tilde{v}_l^{deg} - \frac{\mu}{k_{deg}^l} * \tilde{F}_l = 0$ (MB) _l
$v_l^{tsl} - \sum_j \eta_l^j \cdot v_j^{asm} = 0$	$\tilde{v}_l^{tsl} - \sum_j \eta_l^j \cdot \frac{L_l^{aa}}{k_{cat}^{rib} R_l^{ub}} \cdot k_{deg}^j E_j^{ub} \cdot \tilde{v}_j^{asm} = 0$ (PB) _l
$-v_{aa_i}^{charging} + \sum_l \eta_{aa_i}^l \cdot v_l^{tsl} - \mu * T_{aa_i}^u = 0$	$-\frac{1}{\bar{\mu} T_{aa_i}^{ub}} v_{aa_i}^{charging} + \sum_l \eta_{aa_i}^l \cdot \frac{1}{\bar{\mu} T_{aa_i}^{ub}} \cdot \frac{k^{rib} R_l^{ub}}{L_l^{aa}} \cdot \tilde{v}_l^{tsl} - \frac{\mu}{\bar{\mu}} * \tilde{T}_{aa_i}^u = 0$ (TB) _{aa,i,u}
$v_{aa_i}^{charging} - \sum_l \eta_{aa_i}^l \cdot v_l^{tsl} - \mu * T_{aa_i}^c = 0$	$\frac{1}{\bar{\mu} T_{aa_i}^{ub}} v_{aa_i}^{charging} - \sum_l \eta_{aa_i}^l \cdot \frac{1}{\bar{\mu} T_{aa_i}^{ub}} \cdot \frac{k^{rib} R_l^{ub}}{L_l^{aa}} \cdot \tilde{v}_l^{tsl} - \frac{\mu}{\bar{\mu}} * \tilde{T}_{aa_i}^u = 0$ (TB) _{aa,i,c}
$v_{rRNA_l}^{tcr} - v_{rib}^{asm} = 0$	$\frac{1}{k_{deg}^j E_j^{ub}} \cdot \frac{k_{cat}^{RNAP} P_l^{ub}}{L_l^{nt}} \cdot \tilde{v}_{rRNA_l}^{tcr} - \tilde{v}_{rib}^{asm} = 0$ (RB) _{rRNA,l}
<i>Macromolecule degradation</i>	
$v_j^{deg} - k_{deg}^j \cdot E_j = 0$	$\tilde{v}_j^{deg} - \tilde{E}_j = 0$ (ED) _j
$v_l^{deg} - k_{deg}^l \cdot F_l = 0$	$\tilde{v}_l^{deg} - \tilde{F}_l = 0$ (MD) _l
<i>Transcript./translation catalytic constraints</i>	
$v_l^{tsl} - \frac{k^{rib}}{L_l^{aa}} R_l \leq 0$	$\tilde{v}_l^{tsl} - \tilde{R}_l \leq 0$ (TR2) _l
$v_l^{tcr} - \frac{k_{cat}^{RNAP}}{L_l^{nt}} P_l \leq 0$	$\tilde{v}_l^{tcr} - \tilde{P}_l \leq 0$ (TR1) _l
<i>Expression coupling</i>	
$R_l - \frac{L_l^{nt}}{L_{rib}^{nt}} F_l \leq 0$	$\tilde{R}_l - \frac{L_l^{nt}}{L_{rib}^{nt}} \frac{F_l^{ub}}{R_l^{ub}} \tilde{F}_l \leq 0$ (EX) _l
<i>Capacity constraints</i>	
$R_F - (1 - \rho) E_{rib} = 0$	$\tilde{R}_F - (1 - \rho) \tilde{E}_{rib} = 0$ (RR)
$\sum_l R_l + R_F - E_{rib} = 0$	$\sum_l \frac{R_l^{ub}}{E_{rib}^{ub}} \tilde{R}_l + \tilde{R}_F - \tilde{E}_{rib} = 0$ (TC2)
$\sum_l P_l - E_{RNAP} = 0$	$\sum_l \frac{P_l^{ub}}{E_{RNAP}^{ub}} P_l - \tilde{E}_{RNAP} = 0$ (TC1)
ETFL – Supp. Note 1	2/5

<i>Initial</i>	<i>Scaled</i>
$S \cdot v = 0$	(FBA)
<i>Catalytic constraints</i>	
$v_j^f - k_{cat,f}^j E_j \leq 0$	$\tilde{v}_j^f - \tilde{E}_j \leq 0$ (FC) _j
$v_j^b - k_{cat,b}^j E_j \leq 0$	$\tilde{v}_j^b - \tilde{E}_j \leq 0$ (BC) _j
<i>Macromolecule mass balances</i>	
$v_j^{asm} - v_j^{deg} - \mu * E_j = 0$	$\tilde{v}_j^{asm} - \tilde{v}_j^{deg} - \alpha_j \cdot \mu * \tilde{E}_j = 0$ (EB) _j
$v_l^{tcr} - v_l^{deg} - \mu * F_l = 0$	$\gamma_l \cdot \tilde{v}_l^{tcr} - \tilde{v}_l^{deg} - \beta_j \cdot \mu * \tilde{F}_l = 0$ (MB) _l
$v_l^{tsl} - \sum_j \eta_l^j \cdot v_j^{asm} = 0$	$\tilde{v}_l^{tsl} - \sum_j \eta_l^j \cdot \delta_l^j \cdot \tilde{v}_j^{asm} = 0$ (PB) _l
$-v_{aa_i}^{charging} + \sum_l \eta_{aa_i}^l \cdot v_l^{tsl} - \mu * T_{aa_i}^u = 0$	$-\frac{1}{\bar{\mu} T_{aa_i}^{ub}} v_{aa_i}^{charging} + \sum_l \eta_{aa_i}^l \cdot \tau_l \cdot \tilde{v}_l^{tsl} - \frac{\mu}{\bar{\mu}} * \tilde{T}_{aa_i}^u = 0$ (TB) _{aa_i,u}
$v_{aa_i}^{charging} - \sum_l \eta_{aa_i}^l \cdot v_l^{tsl} - \mu * T_{aa_i}^c = 0$	$\frac{1}{\bar{\mu} T_{aa_i}^{ub}} v_{aa_i}^{charging} - \sum_l \eta_{aa_i}^l \cdot \tau_l \cdot \tilde{v}_l^{tsl} - \frac{\mu}{\bar{\mu}} * \tilde{T}_{aa_i}^c = 0$ (TB) _{aa_i,c}
$v_{rRNA_l}^{tcr} - v_{rib}^{asm} = 0$	$\gamma_l \cdot \tilde{v}_{rRNA_l}^{tcr} - \tilde{v}_{rib}^{asm} = 0$ (RB) _{rRNA,l}
<i>Macromolecule degradation</i>	
$v_j^{deg} - k_{deg}^j \cdot E_j = 0$	$\tilde{v}_j^{deg} - \tilde{E}_j = 0$ (ED) _j
$v_l^{deg} - k_{deg}^l \cdot F_l = 0$	$\tilde{v}_l^{deg} - \tilde{F}_l = 0$ (MD) _l
<i>Transcript./translation catalytic constraints</i>	
$v_l^{tsl} - \frac{k_{cat}^{rib}}{L_{aa}^{nt}} R_l \leq 0$	$\tilde{v}_l^{tsl} - \tilde{R}_l \leq 0$ (TR2) _l
$v_l^{tcr} - \frac{k_{cat}^{RNAP}}{L_l^{nt}} P_l \leq 0$	$\tilde{v}_l^{tcr} - \tilde{P}_l \leq 0$ (TR1) _l
<i>Expression coupling</i>	
$R_l - \frac{L_l^{nt}}{L_{rib}^{nt}} F_l \leq 0$	$\tilde{R}_l - \lambda_l \cdot \tilde{F}_l \leq 0$ (EX) _l
<i>Capacity constraints</i>	
$R_F - (1 - \rho) E_{rib} = 0$	$\tilde{R}_F - (1 - \rho) \tilde{E}_{rib} = 0$ (RR)
$\sum_l R_l + R_F - E_{rib} = 0$	$\sum_l \rho_l \cdot \tilde{R}_l + \tilde{R}_F - \tilde{E}_{rib} = 0$ (TC2)
$\sum_l P_l - E_{RNAP} = 0$	$\sum_l \pi_l \cdot P_l - \tilde{E}_{RNAP} = 0$ (TC1)

Appendix A. ETFL: Supplementary information

<i>Variable</i>	<i>Nondimensionalization factor</i>	<i>Scaled variable</i>
X	α	$\tilde{X} = \frac{X}{\alpha}$
E_j	E_j^{ub}	$\tilde{E}_j = \frac{E_j}{E_j^{ub}}$
F_l	F_l^{ub}	$\tilde{F}_l = \frac{F_l}{F_l^{ub}}$
R_l	R_l^{ub}	$\tilde{R}_l = \frac{R_l}{R_l^{ub}}$
R_F	E_{rib}^{ub}	$\tilde{R}_F = \frac{R_F}{E_{rib}^{ub}}$
P_l	P_l^{ub}	$\tilde{P}_l = \frac{P_l}{P_l^{ub}}$
$T_{aa_i}^u$	$T_{aa_i}^{ub}$	$\tilde{T}_{aa_i}^u = \frac{T_{aa_i}^u}{T_{aa_i}^{ub}}$
$T_{aa_i}^c$	$T_{aa_i}^{ub}$	$\tilde{T}_{aa_i}^c = \frac{T_{aa_i}^c}{T_{aa_i}^{ub}}$
v_j^f	$k_{cat,f}^j E_j^{ub}$	$\tilde{v}_j^f = \frac{v_j^f}{k_{cat,f}^j E_j^{ub}}$
v_j^b	$k_{cat,b}^j E_j^{ub}$	$\tilde{v}_j^b = \frac{v_j^b}{k_{cat,b}^j E_j^{ub}}$
v_j^{deg}	$k_{deg}^j E_j^{ub}$	$\tilde{v}_j^{deg} = \frac{v_j^{deg}}{k_{deg}^j E_j^{ub}}$
v_l^{deg}	$k_{deg}^l F_l^{ub}$	$\tilde{v}_l^{deg} = \frac{v_l^{deg}}{k_{deg}^l F_l^{ub}}$
v_l^{tsl}	$\frac{k_{rib}^{rib}}{L_l^{aa}} R_l^{ub}$	$\tilde{v}_l^{tsl} = \frac{L_l^{aa}}{k_{cat}^{rib} R_l^{ub}} \cdot v_l^{tsl}$
v_l^{tcr}	$\frac{k_{cat}^{RNAP}}{L_l^{nt}} P_l^{ub}$	$\tilde{v}_l^{tcr} = \frac{L_l^{nt}}{k_{cat}^{RNAP} P_l^{ub}} \cdot v_l^{tcr}$
v_j^{asm}	$k_{deg}^j E_j^{ub}$	$\tilde{v}_j^{asm} = \frac{v_j^{asm}}{k_{deg}^j E_j^{ub}}$

<i>Upper bound</i>	<i>(Approximate) Value</i>	<i>Order of magnitude</i>	<i>Unit</i>
E_j^{ub}	$\frac{1}{MW(Enz_j)} \cdot 10^3$	$10^{-2} \cdot 10^3 = 10^1$	$\frac{[g \cdot gDW^{-1}]}{[g \cdot mol^{-1}]} \left[\frac{mmol}{mol} \right]$ $= mmol \cdot gDW^{-1}$
F_l^{ub}	$\frac{1}{MW(mRNA_l)} \cdot 10^3$	$10^{-2} \cdot 10^3 = 10^1$	$mmol \cdot gDW^{-1}$
R_l^{ub}	$\frac{1}{MW(Rib)} \cdot 10^3$	$10^{-6} \cdot 10^3 = 10^{-3}$	$mmol \cdot gDW^{-1}$
P_l^{ub}	$\frac{1}{MW(RNAP)} \cdot 10^3$	$10^{-5} \cdot 10^3 = 10^{-2}$	$mmol \cdot gDW^{-1}$
$T_{aa_i}^{ub}$	$\frac{1}{MW(tRNA_{aa_i})} \cdot 10^3$	$10^{-2} \cdot 10^3 = 10^1$	$mmol \cdot gDW^{-1}$

Nondimensionalization term Expression

α_j	$\frac{\mu}{k_{deg}^j}$
β_l	$\frac{\mu}{k_{deg}^l}$
γ_l	$\frac{1}{k_{deg}^l F_l^{ub}} \cdot \frac{k_{cat}^{RNAP} P_l^{ub}}{L_l^{nt}}$
δ_l^j	$\frac{L_l^{aa}}{k_{cat}^{rib} R_l^{ub}} \cdot k_{deg}^j E_j^{ub}$
$\tau_l^{aa_i}$	$\frac{1}{\bar{\mu} T_{aa_i}^{ub}} \cdot \frac{k^{rib} R_l^{ub}}{L_l^{aa}}$
λ_l	$\frac{L_l^{nt} F_l^{ub}}{L_{rib}^{nt} R_l^{ub}}$
ρ_l	$\frac{R_l^{ub}}{E_{rib}^{ub}}$
π_l	$\frac{P_l^{ub}}{E_{RNAP}^{ub}}$

Supplementary Table 1: Example EP constraint matrix.

	FBA fluxes	biomass reaction flux	rRNA charging reactions	Translation fluxes	Transcription fluxes	Degradation fluxes	Complexation fluxes	Enzyme concentrations	Ribosome concentrations	RNAI concentrations	mRNA concentrations	miRNA concentrations	DNA concentration	Growth discretization variables	product variables	Allocation indicator variables	Allocation variables	61301
FBA mass balances																		1806
Peptide mass balances																		1431
rRNA mass balances																		3
Catalytic constraint																		862
Translation																		1431
Translation capacity																		2
Transcription																		1431
Transcription capacity																		2
Transcription-translation coupling																		1430
Enzyme mass balances																		562
tRNA mass balances																		42
miRNA mass balances																		1431
DNA mass balance																		1
Degradation definition																		1993
Product linearization																		48864
Growth coupling																		1
Growth discretization																		1
AllocationConstraints																		8
	37500	5168	2	42	2862	2862	3986	1124	562	1431	1431	42	1431	1	9	16286	256	3

Standard Operating Procedure to construct an ETFL model.

Summary checklist

Here is a summarized checklist of the material needed to turn a COBRA model into ETFL:

- A working installation of ETFL
- A Cobra model with:
 - Gene identifiers (IDs)
 - All nucleotides triphosphates(NTPs), deoxynucleotides triphosphate(dNTP), nucleotides monophosphate (NMP), aminoacids.
 - (Optional) Gene reaction rules
- Gene sequences indexed by their gene IDs
- Peptide stoichiometry of enzymes
- Enzyme assignments per reaction.
- Enzyme catalytic rate constants:
 - Forward
 - (Optional) Reverse
- Enzyme degradation rate constants
- mRNA degradation rate constants
- (Optional) Free ribosomes ratio
- (Optional) Free RNA Polymerase ratio
- (Optional) GC-content and length of the genome
- (Optional) Average aminoacid abundances
- (Optional) Average NTP abundances
- (Optional) Average mRNA length
- (Optional) Average peptide length
- (Optional) Growth-dependant mRNA, peptide, and DNA mass ratios.

Setup

Prerequisites

Make sure you have `Git` installed. Since ETFL is built upon `pyTFA` [1], we will clone both repositories. In a folder of your choice, download the source code from our repositories:

```
git clone https://github.com/EPFL-LCSB/pytfa
git clone https://github.com/EPFL-LCSB/etfl
# -- OR --
git clone https://gitlab.com/EPFL-LCSB/pytfa
git clone https://gitlab.com/EPFL-LCSB/etfl
```

Docker container (recommended)

We recommend the use of Docker containers as they provide a standardized, controlled and reproducible environment. The ETFL Docker is built upon the `pyTFA` Docker image. We recommend building it yourself as it is where your solvers can be installed.

Downloading Docker

If Docker is not yet installed on your machine, you can get it from [here](#)

Building and running the Docker container

```
# Build the pyTFA docker
cd pytfa/docker && . build
# Build and run the ETFL docker
cd ../../etfl/docker
. build
. run
```

Solvers

For installing the solvers, please refer to the `pyTFA` documentation

Python environment

Alternatively, you can install ETFL using `pip`:

```
pip install etfl
```

Make sure your solvers are also installed in the same environment if you are using a `virtualenv` or `pyenv`.

From COBRA to ETFL

ETFL models can be generated fairly easily from a COBRA model. In the following subsections, we detail the required information to add expression constraints to a COBRA model and turn it into an ETFL model.

Constraint-based model

You will need to start with a COBRA model including the following information:

- Genes and their gene ID (necessary to retrieve gene sequences)
- (Optional) Gene-protein rules: These are used to make approximated enzymes if peptide information is not enough

Additionally, you will need to build a dictionary of essential metabolites required in the model. It should follow this example structure (all fields mandatory):

```
dict(atp='atp_c',
     adp='adp_c',
     amp='amp_c',
     gtp='gtp_c',
     gdp='gdp_c',
     pi='pi_c',
     ppi='ppi_c',
     h2o='h2o_c',
     h='h_c')
```

A dictionary of RNA NTPs, DNA dNTPS, and aminoacids is also required, of the type:

```
aa_dict = {'A': 'ala_L_c',
           # ...
           'V': 'val_L_c', }

rna_nucleotides = {
    'u': 'utp_c',
    # ...
    'c': 'ctp_c'}

rna_nucleotides_mp = {
    'u': 'ump_c',
    # ...
    'c': 'cmp_c'}

dna_nucleotides = {
    't': 'dttp_c',
    # ...
    'c': 'dctp_c'}
```

From genes to peptides

In order to build the transcription and translation, it is necessary to provide ETFL with gene deoxynucleotide sequences. These will be automatically transcribed in RNA sequences and then translated into aminoacid peptide sequences. They must be fed to the function `model.add_nucleotides_sequences` in a

dict-like object, indexed by gene IDs (`model.genes.mygene.id` property in COBRA).

We suggest the following sources for obtaining such information:

- KEGG Genes
- NCBI Gene DB
- MetaCyc Gene Search

ETFL will automatically synthesize the correct peptides from the nucleotides sequences. This is based on the Biopython package's `transcribe` and `translate` functions [2].

For each enzyme created by transcription, a degradation rate constant must be specified. These can be obtained through literature search, or using an average value.

From peptides to enzymes

A key part of the expression modeling is to properly represent the assembly of enzymes from peptides. For each enzyme of the model, a stoichiometry of the peptides necessary for its assembly is needed. These are stored as dictionaries in the `Enzyme.composition` property under a form similar to :

```
>>> enzyme.composition
{'b2868': 1, 'b2866': 1, 'b2867': 1}
```

The keys match the IDs of genes coding for the peptide, and the value represent the stoichiometry of the peptide in the enzyme. These can be obtained from literature search or specialized databases. In particular, we used for this paper the Metacyc/Biocyc database [3, 4], using specialised SmartTables queries [5].

```
html-sort-ascending(
  html-table-headers (
    [(f,genes,(protein-to-components f)):
      f<-ECOLI^Protein-Complexes,genes := (enzyme-to-genes f)
    ],
    ("Product Name", "Genes", "Component coefficients")),
  1)
```

At this step, it is also possible to implement post-translational changes or enzyme-specific mechanisms. The assembly reaction of peptides can be edited like any normal reaction to include other metabolites, for example metal ions.

From enzymes back to the metabolism

Lastly, the enzymes must be assigned reactions and catalytic rate constants. Several enzymes can catalyze the same reactions. COBRA models can take this into account differently, usually having either (i) multiple reactions with a simple gene reaction rule; or (ii) one unique reaction with several isozymes in the gene reaction rule. Although not often applied consistently within the same

model, these two formalisms are equivalent, and their ETFL counterparts will also behave equivalently.

For each enzyme, the information needed is the (forward) catalytic rate constant k_{cat}^+ , facultatively the reverse catalytic rate constant k_{cat}^- (set equal to k_{cat}^+ if none is given), and a degradation rate constant.

This is done by calling the function `model.add_enzymatic_coupling(coupling_dict)` where `coupling_dict` is a dict-like object with reaction IDs as keys and a list of enzyme objects as values:

```
coupling_dict = {
    #...
    'AB6PGH': [ <Enzyme AB6PGH_G495_MONOMER at 0x7ff00e0f1b38>],
    'ABTA'  : [ <Enzyme ABTA_GABATRANSAM at 0x7ff00e0fda90>,
               <Enzyme ABTA_G6646 at 0x7ff00e0fd4e0>],
    'ACALD' : [ <Enzyme ACALD_MHPF at 0x7ff00e0fdcf8>],
    #...
}
```

The catalytic rate constants can be obtained from several databases, such as:

- Rhea
- BRENDA
- SabioRK
- Uniprot

Several enzymes can be assigned to a reaction. ETFL will try to match the gene reaction rule isozymes to the supplied enzymes. If the gene reaction rule shows several isozymes while only one enzyme is supplied, the enzyme can be replicated to match the number of isozymes in the gene reaction rule.

Given a reaction in the model, if no enzyme is supplied but the reaction possesses a gene reaction rule, it is possible to infer an enzyme from it. The rule expression is expanded, and each term separated by an **OR** boolean operator is interpreted as an isozyme, while terms separated by an **AND** boolean operators are interpreted as unit peptide stoichiometric requirements. The enzyme is then assigned an average catalytic rate constant and degradation rate constant.

Growth-dependant parameters

Accounting for growth-dependent RNA and protein content requires additional information. In particular:

- GC-content and length of the genome
- Average aminoacid abundances
- Average NTP abundances
- Average mRNA length

- Average peptide length
- Growth-dependant mRNA, peptide, and DNA mass ratios.

These values are usually obtained through literature search. All of the last three ratios are optional, although using none defeats the purpose of accounting for growth-dependant parameters.

Additional documentation

Example

We encourage the reader to look at the script used to generate the models with which the paper's results were generated, available in `etfl/tutorials/helper_gen_models.py`. The data it takes in input has been generated in `etfl/etfl/data/ecoli.py`. These are good examples to start from in order to make a custom ETFL from a different COBRA model.

Note on steady-state assumptions and dilution terms.

Flux balance analysis (FBA) is an important tool in metabolic engineering to analyze the stoichiometric properties of living systems. Its success is partly due to the simplicity of its formulation as a linear program, with a constraint matrix of the form $S \cdot v = 0$, where S is the stoichiometric matrix of the system of interest, and v the biochemical fluxes carried by the reactions in the system. This formulation is directly derived from the mass balance of the metabolites inside the cell. The simplicity of this formulation stems from two important assumptions: (i) in the mass balance of the metabolites, the dilution term is negligible ; and (ii) the intracellular concentrations of the metabolites are at quasi-steady state.

The main purpose of ME-models is to account for macromolecule synthesis costs on top of a metabolic model. However, the macromolecular concentrations are subject to different assumptions. In particular, when writing the mass balances for the said macromolecules, the dilution term is not negligible anymore. Furthermore, the quasi-steady state assumption applies on a different timescale, since macromolecules synthesis rates are several orders of magnitude slower than metabolic reactions.

Here we present three arguments to explain and justify the assumptions made in ETFL and the form of the mass balance equations for metabolites and macromolecules. We briefly discuss these arguments for the case of metabolites and contrast them in the case of macromolecules, to study the validity of the assumptions made in ETFL.

Preliminaries

The mass balances of biochemical species is written with respect to their concentration variables. If we assume the cell is growing at a specific growth rate μ , we must assume that the volume of cell within which the mass balance is considered varies.

The mass balance of a compound X can be expressed both as the derivative of the mass or the algebraic sum of its synthesis and consumption fluxes:

$$\frac{dm_X}{dt} = C_X \frac{dV_c}{dt} + V_c \frac{dC_X}{dt} \quad (1)$$

$$= S_X^\top \cdot v \cdot V_c, \quad (2)$$

where C_X is the concentration of compound X in the cellular volume V_c , for a total mass m_X in the cell, and whose stoichiometry with respect to the fluxes v is described by the row S_X of the stoichiometric matrix S .

We next combine equations 1 and 2 and divide by V_c (necessarily non-zero) to write the time derivative of the concentration C_X :

$$\frac{dC_X}{dt} = S_X^\top \cdot v - \frac{1}{V_c} \frac{dV_c}{dt} \cdot C_X. \quad (3)$$

By definition, $\frac{1}{V_c} \frac{dV_c}{dt} = \mu$ is the specific growth rate of the cell (under the assumption of constant cell density ρ_c), and the term $\mu \cdot C_X$ is called the dilution term, as per Fredrickson's work on formulating growth models [6]. We can hence write the general mass balance of a biochemical species in the cell as:

$$\frac{dC_X}{dt} = S_X^\top \cdot v - \mu \cdot C_X. \quad (4)$$

In this equation, rates are in $\text{g}/(\text{L} \cdot \text{h})$, and concentrations in g/L . If we divide rates and concentrations by their respective molecular weight and the mass of one liter of dried cells, their units become respectively $\text{mmol}/(\text{gDW} \cdot \text{h})$, and mmol/gDW . We will use the latter unit system in the rest of this note.

Intracellular fluxes and dilution

In FBA, the dilution term is omitted from the mass balance of the metabolites. In ETFL, this term is also omitted for metabolites, but preserved for macromolecules. We present here two arguments which support the fact that the contribution of the dilution is negligible for metabolites, but not for macromolecules.

Orders of magnitude argument

The average metabolite concentration in the cells do not exceed $10^{-2} \text{ M} = 10 \text{ mmol/L}$ [7]. Assuming the cell has a density close to $1 \text{ kg/L} = 1000 \text{ g/L}$, and that $0.5 \approx 10^0 \text{ gDW/g}$ of dry cells is obtained per gram of culture, we derive that:

Intracellular metabolite concentrations are upperbounded by 10^{-2} mmol/gDW .

From typical FBA results and flux variability analyses, we can claim the following:

Metabolic fluxes typically range from 10^{-2} to $10^1 \text{ mmol}/(\text{gDW} \cdot \text{h})$.

These fluxes are higher close to the carbon uptake, in the central carbon metabolism, and decrease in the more distant pathways.

One *E. coli* cell weighs 1 pg . Using the previous constants yields a conversion factor of $10^{-8} (\text{mmol}_{\text{cell}})/\text{gDW}$. The number of mRNA copies per cell per transcript is in the order of magnitude 10^0 copies/cell (BNID 112795 [8]). This amounts to a typical mRNA concentration of $10^{-8} \text{ mmol}_{\text{mRNA}}/\text{gDW}$. Protein-to-mRNA ratios are typically ranging from 10^2 to $10^4 \text{ proteins/mRNA}$ (BNID 106254 [9]).

From this we can assert:

Intracellular macromolecule concentrations range from 10^{-8} to 10^{-4} mmol/gDW .

There are in average 6.6 ribosomes per thousand base pairs per cell (BNID 107727 [10]), and the average transcript is around 1 kb , with one copy per transcript, which amounts to $10^1 \text{ mmol}_{\text{rib}}/(\text{cell} \cdot \text{transcript})$. The translation rate per ribosome is $10 \text{ aa}/(\text{s} \cdot \text{ribosome})$ (BioNumbers ID [BNID] 100059 [11]). This gives an upper bound on the specific peptide synthesis fluxes of $v^{tsl} \approx$

10^{-6} mmol/(gDW · h) Using the protein-to-mRNA ratio allows us to estimate an upperbound on transcription rates from 10^{-10} to 10^{-8} mmol/(gDW · h). This yields:

The typical macromolecule synthesis rate range from 10^{-10} to 10^{-6} mmol/(gDW · h).

An interesting intermediary case to consider is that of macromolecule monomers (nucleotides for mRNA and amino acids for peptides). Under the assumption that the typical protein is ≈ 325 amino acids long (BNID 108986 [12]), the average mRNA transcript is $\approx 1kb$, and there are $\approx 10^3$ different mRNAs and peptides, we can derive typical monomer concentrations for each of the ≈ 20 amino acids and 4 nucleotides. Thus, nucleotides have a typical concentration of 10^{-3} mmol_{nt}/gDW, and amino acids have a typical concentration of 10^{-3} to 10^{-1} mmol_{aa}/gDW. We can then assert:

Macromolecule monomers have a typical concentration of 10^{-3} to 10^{-1} mmol/gDW.

Assuming that 50% of the glucose goes towards protein synthesis, and that all the ≈ 20 amino acids are synthesized in similar amounts at yields between 0.5 and $2.0 \text{ mol}_{aa}/\text{mol}_{glc}$ [13], the amino acid biosynthesis fluxes are one to two orders of magnitude smaller than those of the central carbon metabolism. Nucleotide synthesis is even smaller. From there, we can claim:

Monomer synthesis fluxes range from 10^{-2} to 10^{-1} mmol/(gDW · h)

The values we obtain for the elements of Eq. 4 are detailed in Table 3. The table shows clearly that, in the case of metabolites, the dilution term is negligible in front of the metabolic fluxes. It also shows that for macromolecule monomers, which are further away from the central carbon metabolism, the dilution term becomes comparable with the synthesis term. These orders of magnitude are in agreement with the comprehensive discussion on the magnitude of pools, metabolic fluxes, and dilution terms for different metabolites in the cell (including central carbon pathway and amino acids) featured in the chapter 8.1 of the work by Stephanopoulos, Aristidou and Nielsen [14]. Finally, the range of synthesis fluxes for macromolecules greatly overlaps with that of their dilution term, which imposes taking the dilution into account.

Yield argument

We have seen before that the average amino acid yield per molecule of glucose metabolised is between 0.5 and $2.0 \text{ mol}_{aa}/\text{mol}_{glc}$ [13]. The average protein is

Supplementary Table 2: Orders of magnitudes of the variables in presence for the mass balance of metabolites.

Variable	Order of magnitude			Units
	Metabolites	Monomers	Macromolecules	
$S_X^T \cdot v$	$10^{-2} - 10^1$	$10^{-2} - 10^{-1}$	$10^{-10} - 10^{-6}$	mmol/(gDW · h)
$\mu \cdot C_X$	$10^{-3} - 10^{-2}$	$10^{-4} - 10^{-1}$	$10^{-9} - 10^{-4}$	mmol/(gDW · h)

made of ≈ 325 amino acids (BNID 108986 [12]). This implies that the global protein synthesis rate is at least 2 orders of magnitude slower than those of the central carbon metabolism. Given approximately 10^3 different peptides, one can expect specific peptide synthesis rates to be at their maximum 5 orders of magnitude smaller than central carbon metabolism fluxes.

This estimation matches with the results presented in Table 3.

An additional note about accounting for dilution in FBA

Benyamini *et al.* [15] report on the results of FBA accounting for dilution terms, a method called MD-FBA. Their results show a sensitivity in gene essentiality analysis, especially for genes far from the central carbon pathways, and closer to the biomass precursor pathways. In particular, the fluxes in which the dilution term has a significant impact are the fluxes close to the synthesis of the macromolecule monomers.

This can be understood in the context of the order-of-magnitude argument since breaking up the glucose to piece together biomass precursors further splits the available carbon between the different precursors, thus reducing their synthesis fluxes, which then become comparable to the dilution rate.

Taken to the extreme, this reasoning matches the yield argument made previously, where specific peptide synthesis fluxes will be several orders of magnitude smaller than the central metabolism fluxes, making the dilution term non-negligible.

Timescale analysis

Heijnen *et al.*'s analysis of the pseudo-steady state hypothesis for biochemical kinetics [16] provides a method to study kinetic equations using non-dimensionalization. We adapt this method to Eq. 4 to derive a justification of the pseudo-steady state hypothesis.

Let us assume that the general flux term $v(t)$ can be written as the product of a diagonal matrix of catalytic rate constants K and a function of concentration of the compounds taking part in the reactions $\Phi(C, t)$, with $C(t) = (C_X(t))_{X \in \text{species}}$. This product can represent, for instance, either mass action kinetics, or Michaelis-Menten kinetics. We can hence rewrite Eq. 4:

$$\frac{dC_X(t)}{dt} = S_X^\top \cdot K \cdot \Phi(C, t) - \mu \cdot C_X(t). \quad (5)$$

In a Michaelis-Menten case, for example, the elements of the matrix K will represent the catalytic rate constants k_{cat} .

Quasi-steady state assumption

We introduce the dimensionless variables:

$$\tau = kt, \quad z(\tau) = \frac{C_X(\tau/k)}{c_0}, \quad y(\tau) = \frac{\Phi(C, \tau/k)}{c_0}, \quad (6)$$

where k is an inverse time constant of our choice and c_0 is the average species concentration, acting as a non-dimensionalization factor. We can rewrite Eq. 5:

$$\frac{dz}{d\tau} = S_X^\top \cdot \left(\frac{1}{k} K \right) \cdot y(\tau) - \frac{\mu}{k} \cdot z(\tau) = F(y, z, \tau). \quad (7)$$

In Heineken's words, if k is sufficiently large, τ represents time vastly accelerated so that $F(y, z, \tau)$ is held at a stable root of $F(y, z, \tau) = 0$ (which exists if $\Phi(C, t)$ is sufficiently well-behaved, by following for instance Michaelis-Menten kinetics [16]). This is the quasi-steady state assumption.

We must compare our choice of k to the inverse characteristic time of change in physiology of our cells. Let us note the latter is not μ , but rather the characteristic time of change of the experimental properties, such as the concentration of species in the culture medium. In an ideal continuous culture, this time should be the whole observation time of the exponential growth, since cell physiology in ideal culture conditions should not change.

The slowest $k = k_{slow}$ we can choose is the ribosome transcription rate for the average peptide. We have:

$$k = k_{slow} = \frac{k_{trans}}{L_{avg}^{aa}} \approx \frac{10}{325} * 3600 \approx 10^2 \text{ h}^{-1}. \quad (8)$$

This yields the characteristic time $t_{slow} = 1/k_{slow} = 10^{-2} \text{ h} \approx 30 \text{ s}$. As long as the characteristic time of change in physiology is longer than the slowest mode of our system, the steady state assumption holds. Since k_{slow} lowerbounds all the other rate constants of the system, if the steady state assumption is valid for k_{slow} , then the steady-state assumption is valid for all the other parameters and variables of the system. We can then formulate the steady-state assumption:

As long as the characteristic time of change of the physiology is much longer than 30 s, the steady-state assumption is valid for all variables and parameters of the model.

Dilution rate

We can also use Eq. 7 to study the contribution of each term of the right-hand side of the equation.

In a Michaelis-Menten case, we set $k = k_{cat}$ and rewrite:

$$\left(\frac{1}{k} K \right) \cdot y = \frac{1}{k_{cat}} \cdot \frac{k_{cat}}{c_0} \cdot \Phi(C), \quad (9)$$

$$= \frac{1}{k_{cat} \cdot e_0} \cdot \frac{e_0}{c_0} \cdot k_{cat} \cdot \Phi(C), \quad (10)$$

$$= \frac{e_0}{c_0} \cdot \frac{v}{V_{max}} \quad (11)$$

with K_M, V_{max} the usual Michaelis-Menten constants, v the flux we are studying, and we chose $e_0 = E_{qss}$ the total concentration of the enzyme catalyzing the reaction at quasi-steady state. Since $v/V_{max} \approx 1$ in terms of orders of magnitude, and $z = C_X/c_0 \approx 1$ also, comparing the two right-hand side terms of Eq. 7 is equivalent to comparing e_0/c_0 and μ/k .

In particular, if we consider X to be a metabolite, we can set $k \approx 10^2 \text{ s}^{-1} \approx 10^5 \text{ h}^{-1}$, a typical catalytic rate constant for metabolic reactions, and $c_0 \approx 10^{-2} \text{ mmol/gDW}$ and $e_0 \approx 10^{-5} \text{ mmol/gDW}$ according to section , we obtain:

$$\frac{e_0}{c_0} \approx 10^{-3}, \quad \frac{\mu}{k} \approx 10^{-5}. \quad (12)$$

The dilution term appears to be negligible for metabolites.

In the case of peptides, we must set $k = k_{tsl} \approx 10^2 \text{ h}^{-1}$ (see Eq. 8). The substrates of the translation are amino acids, hence $c_0 \approx 10^{-2} \text{ mmol}_{aa}/\text{gDW}$ and the ribosome concentration is $e_0 \approx 10^{-4} \text{ mmol}/\text{gDW}$ according to section . This yields:

$$\frac{e_0}{c_0} \approx 10^{-2}, \quad \frac{\mu}{k} \approx 10^{-2}. \quad (13)$$

The dilution term appears to be non-negligible for peptides. A similar reasoning can be performed for mRNA synthesis.

With this argument, we can recover the argument on the orders of magnitude made in section , since comparing e_0/c_0 and μ/k is equivalent to evaluating the following quotient:

$$q = \frac{e_0/c_0}{\mu/k} \approx \frac{V_{max}}{\mu \cdot C_X}, \quad (14)$$

If $q \gg 1$, then the dilution term is negligible in front of the metabolic fluxes. If $q \ll 1$, then the dilution term is preponderant. Finally, if $q \approx 1$, then both terms need to be taken into account. We previously showed that, for metabolites, the dilution term is almost always negligible in front of the metabolic fluxes, a result which we recover in Eq. 12, where $q \gg 1$. We also showed previously that macromolecule dilutions are not always negligible in front of their synthesis rates. We recovered this result for peptides, for which we evaluate $q \approx 1$ in Eq. 13.

Balanced growth hypothesis

The idea of balanced growth was introduced by Monod [17], and has been refined further by Campbell [18]. Monod's approach, based on Hinshelwood's work [19], explains that once past the lag phase, the cells reach a stable enzyme composition:

“[...] the lag and acceleration phases represent essentially a process of equilibration, the functioning of a regulatory mechanism, by virtue of which a certain enzyme balance inside the cells is attained.”

Monod J., *The growth of bacterial cultures*

Campbell refines this definition in his work from 1957 [18] by the following statement:

“[...] growth is balanced over a time interval if, during that interval, every extensive property of the growing system increases by the same factor.”

Campbell A., *Synchronization of cell division*

Using this definition, and accounting for the fact that species concentrations inside the cells are the quotient of the extensive factors mass (of the said species) and volume (of the cell), we obtain directly that intracellular concentrations are constant under the hypothesis of balanced growth. Campbell adds that this approximation is well suited for a continuous culture, and is well approximated in a batch reactor.

ETFL Formulation

Conventions

Supplementary Table 3: Indices used in the formulation.

Index letter	Indexed variables	Indexing set
i	Metabolite	\mathcal{I}
aa_i	Amino acid	\mathcal{A}
j	Reaction/Flux/Enzyme	\mathcal{J}
l	Gene/Peptide/mRNA	\mathcal{L}
s	Binary coefficient for growth discretization	$\mathcal{S} = \{0..\lceil \log_2 N \rceil\}$
u	Binary coefficient for interpolation discretization	$\mathcal{U} = \{0..N\}$

Supplementary Table 4: Variables used in the formulation.

Symbol	Variable	Unit
μ	Growth rate	h^{-1}
v_j^\pm	j^{th} net positive/negative biochemical flux	$\text{mmol.gDW}^{-1}.\text{h}^{-1}$
E_j	Concentration of the j^{th} enzyme	mmol.gDW^{-1}
F_l	Concentration of the l^{th} mRNA	mmol.gDW^{-1}
P_l	Concentration of the RNA polymerase assigned to the l^{th} mRNA	mmol.gDW^{-1}
R_l	Concentration of the ribosome assigned to the l^{th} peptide	mmol.gDW^{-1}
$T_{aa_i}^u$	Concentration of the i^{th} uncharged tRNA	mmol.gDW^{-1}
$T_{aa_i}^c$	Concentration of the i^{th} charged tRNA	mmol.gDW^{-1}
v_l^{sl}	Translation rate of the l^{th} gene	$\text{mmol.gDW}^{-1}.\text{h}^{-1}$
v_l^{tcr}	Transcription rate of the l^{th} gene	$\text{mmol.gDW}^{-1}.\text{h}^{-1}$
v_j^{asm}	Assembly rate of the j^{th} enzyme	$\text{mmol.gDW}^{-1}.\text{h}^{-1}$
v_j^{deg}	Degradation rate of the j^{th} enzyme	$\text{mmol.gDW}^{-1}.\text{h}^{-1}$
v_l^{deg}	Degradation rate of the l^{th} mRNA	$\text{mmol.gDW}^{-1}.\text{h}^{-1}$
$v_{aa_i}^{\text{charging}}$	Charging rate of the i^{th} tRNA	$\text{mmol.gDW}^{-1}.\text{h}^{-1}$

Supplementary Table 5: Parameters used in the formulation.

Symbol	Parameter	Unit
$k_{\text{cat}}^{j,\pm}$	Forward/backward catalytic rate constant of the j^{th} net biochemical flux	h^{-1}
k_{deg}^j	Degradation rate constant of the j^{th} enzyme	h^{-1}
k_{deg}^l	Degradation rate constant of the l^{th} mRNA	h^{-1}
η_l^j	Stoichiometry of the l^{th} peptide in the j^{th} enzyme	$[\emptyset]$
$\eta_{\text{aa}_i}^l$	Stoichiometry of the i^{th} amino acid in the l^{th} peptide	$[\emptyset]$
L_l^{aa}	Length in amino acids (aa) of the l^{th} peptide	aa
L_l^{nt}	Length in nucleotides (nt) of the l^{th} mRNA	b
$L_{\text{rib}}^{\text{nt}}$	Ribosome footprint size on mRNA, in nucleotides	b
ρ	Ribosome occupancy	$[\emptyset]$
π	RNA polymerase occupancy	$[\emptyset]$

Bilinear formulation

$$\begin{aligned}
& \underset{\mu, v, E, F, R, P, T}{\text{maximize}} && \mu \\
& \text{subject to} && S \cdot v = 0, \\
& && v_j^+ - k_{\text{cat}}^{j,+} E_j \leq 0, \quad \forall j \in \mathcal{J}, \\
& && v_j^- - k_{\text{cat}}^{j,-} E_j \leq 0, \quad \forall j \in \mathcal{J}, \\
& && v_l^{\text{tsl}} - \sum_{j \in \mathcal{J}} \eta_l^j \cdot v_j^{\text{asm}} = 0, \quad \forall l \in \mathcal{L}, \\
& && v_{\text{rRNA}_l}^{\text{tr}} - v_{\text{rib}}^{\text{asm}} = 0, \quad \forall l \in \mathcal{L}, \\
& && v_j^{\text{asm}} - v_j^{\text{deg}} - \mu * E_j = 0, \quad \forall j \in \mathcal{J}, \\
& && v_l^{\text{tr}} - v_l^{\text{deg}} - \mu * F_l = 0, \quad \forall l \in \mathcal{L}, \\
& && -v_{\text{aa}_i}^{\text{charging}} + \sum_{l \in \mathcal{L}} \eta_{\text{aa}_i}^l \cdot v_l^{\text{tsl}} - \mu * T_{\text{aa}_i}^u = 0, \quad \forall \text{aa}_i \in \mathcal{A}, \\
& && v_{\text{aa}_i}^{\text{charging}} - \sum_{l \in \mathcal{L}} \eta_{\text{aa}_i}^l \cdot v_l^{\text{tsl}} - \mu * T_{\text{aa}_i}^c = 0, \quad \forall \text{aa}_i \in \mathcal{A}, \\
& && v_j^{\text{deg}} - k_{\text{deg}}^j \cdot E_j = 0, \quad \forall j \in \mathcal{J}, \\
& && v_l^{\text{deg}} - k_{\text{deg}}^l \cdot F_l = 0, \quad \forall l \in \mathcal{L}, \\
& && v_l^{\text{tr}} - \frac{k_{\text{cat}}^{\text{RNAP}}}{L_l^{\text{nt}}} P_l \leq 0, \quad \forall l \in \mathcal{L}, \\
& && v_l^{\text{tsl}} - \frac{k_{\text{cat}}^{\text{rib}}}{L_{\text{aa}}^{\text{aa}}} R_l \leq 0, \quad \forall l \in \mathcal{L}, \\
& && R_l - \frac{L_l^{\text{nt}}}{L_{\text{rib}}^{\text{nt}}} F_l \leq 0, \quad \forall l \in \mathcal{L}, \\
& && \sum_{l \in \mathcal{L}} R_l + R_{\text{F}} - E_{\text{rib}} = 0, \\
& && \sum_{l \in \mathcal{L}} P_l + P_{\text{F}} - E_{\text{RNAP}} = 0, \\
& && R_{\text{F}} - (1 - \rho) E_{\text{rib}} = 0, \\
& && P_{\text{F}} - (1 - \pi) E_{\text{RNAP}} = 0.
\end{aligned} \tag{1}$$

Integer-linearized formulation

$$\begin{aligned}
 & \underset{\mu, v, E, F, R, P, T}{\text{maximize}} && \mu \\
 & \text{subject to} && \\
 & S \cdot v = 0, \\
 & v_j^+ - k_{\text{cat}}^{j,+} E_j \leq 0, \quad \forall j \in \mathcal{J}, \\
 & v_j^- - k_{\text{cat}}^{j,-} E_j \leq 0, \quad \forall j \in \mathcal{J}, \\
 & v_l^{\text{tsl}} - \sum_{j \in \mathcal{J}} \eta_l^j \cdot v_j^{\text{asm}} = 0, \quad \forall l \in \mathcal{L}, \\
 & v_{\text{rRNA}_l}^{\text{tr}} - v_{\text{rib}}^{\text{asm}} = 0, \quad \forall l \in \mathcal{L}, \\
 & v_j^{\text{asm}} - v_j^{\text{deg}} - \mu * E_j = 0, \quad \forall j \in \mathcal{J}, \\
 & v_l^{\text{tr}} - v_l^{\text{deg}} - \mu * F_l = 0, \quad \forall l \in \mathcal{L}, \\
 & -v_{\text{aa}_i}^{\text{charging}} + \sum_{l \in \mathcal{L}} \eta_{\text{aa}_i}^l \cdot v_l^{\text{tsl}} - \mu * T_{\text{aa}_i}^u = 0, \quad \forall \text{aa}_i \in \mathcal{A}, \\
 & v_{\text{aa}_i}^{\text{charging}} - \sum_{l \in \mathcal{L}} \eta_{\text{aa}_i}^l \cdot v_l^{\text{tsl}} - \mu * T_{\text{aa}_i}^c = 0, \quad \forall \text{aa}_i \in \mathcal{A}, \\
 & v_j^{\text{deg}} - k_{\text{deg}}^j \cdot E_j = 0, \quad \forall j \in \mathcal{J}, \\
 & v_l^{\text{deg}} - k_{\text{deg}}^l \cdot F_l = 0, \quad \forall l \in \mathcal{L}, \\
 & v_l^{\text{tr}} - \frac{k_{\text{cat}}^{\text{RNAP}}}{L_l^{\text{nt}}} P_l \leq 0, \quad \forall l \in \mathcal{L}, \quad (2) \\
 & v_l^{\text{tsl}} - \frac{k_{\text{cat}}^{\text{rib}}}{L_l^{\text{aa}}} R_l \leq 0, \quad \forall l \in \mathcal{L}, \\
 & R_l - \frac{L_l^{\text{nt}}}{L_l^{\text{rib}}} F_l \leq 0, \quad \forall l \in \mathcal{L}, \\
 & \sum_{l \in \mathcal{L}} R_l + R_{\text{F}} - E_{\text{rib}} = 0, \\
 & \sum_{l \in \mathcal{L}} P_l + P_{\text{F}} - E_{\text{RNAP}} = 0, \\
 & R_{\text{F}} - (1 - \rho) E_{\text{rib}} = 0, \\
 & P_{\text{F}} - (1 - \pi) E_{\text{RNAP}} = 0, \\
 & \sum_{j \in \mathcal{J}} \text{MW}_j \cdot E_j - \sum_{u \in \mathcal{U}} \lambda_u \cdot P_u^m = 0, \\
 & \sum_{l \in \mathcal{L}} \text{MW}_l \cdot F_l - \sum_{u \in \mathcal{U}} \lambda_u \cdot R_u^m = 0, \\
 & \text{MW}_{\text{DNA}} \cdot \text{DNA} - \sum_{u \in \mathcal{U}} \lambda_u \cdot \text{Dm}_u = 0, \\
 & \text{[see next page]}
 \end{aligned}$$

$$\begin{aligned}
& \underset{\mu, v, E, F, R, P, T}{\text{maximize}} && \mu \\
& \text{subject to} && [contd.] \quad , \\
& && \sum_{s=0}^{\lceil \log_2 N \rceil} 2^s \cdot \delta_s \leq N \quad , \\
& && \mu - \hat{\mu} \leq \frac{p}{N} \quad , \\
& && \hat{\mu} - \mu \leq \frac{p}{N} \quad , \\
& && \sum_{u \in \mathcal{U}} \lambda_u = 1 \quad , \\
& \sum_{u \in \mathcal{U}} u \cdot \lambda_u - \sum_{l \in \mathcal{L}} 2^l \cdot \delta_l = 0 \quad , \\
& E_j + M \cdot \delta_s - z_j^s \leq M, \quad \forall j \in \mathcal{J}, \\
& z_j^s - M \cdot \delta_s \leq 0 \quad , \quad \forall j \in \mathcal{J}, \\
& z_j^s - E_j \leq 0 \quad , \quad \forall j \in \mathcal{J}, \\
& F_l + M \cdot \delta_s - z_l^s \leq M, \quad \forall l \in \mathcal{L}, \\
& z_l^s - M \cdot \delta_s \leq 0 \quad , \quad \forall l \in \mathcal{L}, \\
& z_l^s - F_l \leq 0 \quad , \quad \forall l \in \mathcal{L}, \\
& T_{aa_i}^u + M \cdot \delta_s - z_{aa_i}^{u,s} \leq M, \quad \forall aa_i \in \mathcal{A}, \\
& z_{aa_i}^{u,s} - M \cdot \delta_s \leq 0 \quad , \quad \forall aa_i \in \mathcal{A}, \\
& z_{aa_i}^{u,s} - T_{aa_i}^u \leq 0 \quad , \quad \forall aa_i \in \mathcal{A}, \\
& T_{aa_i}^c + M \cdot \delta_s - z_{aa_i}^{c,s} \leq M, \quad \forall aa_i \in \mathcal{A}, \\
& z_{aa_i}^{c,s} - M \cdot \delta_s \leq 0 \quad , \quad \forall aa_i \in \mathcal{A}, \\
& z_{aa_i}^{c,s} - T_{aa_i}^c \leq 0 \quad , \quad \forall aa_i \in \mathcal{A}
\end{aligned} \tag{3}$$

ETFL Glossary

Big-M value A value that is systematically bigger than the other variables in presence within an expression. Used with binary variables to model **if**-type logical dependencies in an optimization problem. Often annotated M in expressions.

Bilinear(ity) A function is said to be bilinear if it contains a product of two of its variables. This term is called a bilinearity. A problem with a constraint defined by a bilinear function of variables is said to be bilinear. That is the case in the non-linearized expression problem with the term $\mu * E_j$, where both μ and E_j are variables of the problem.

Binary variable An integer variable whose value is constrained to 0 or 1. Used to model **if**-type logical dependencies in an optimization problem. For instance, they are used in TFA to enforce the statement “if the Gibbs free energys of this reaction is negative, its net biochemical flux will be in its forward direction”. Inclusion of binary variables in a LP problem make it MILP.

Discretization Process by which a continuous variable is replaced by a set of representative discrete values it can take. We use it in ETFL to approximate μ and perform a linearization. Sampling is a type of discretization.

Linearization Process by which a non-linear function is approximated by a linear approximant. In the case of ETFL, we discretize μ to make the bilinear terms $\mu * E_j$ (piecewise-)linear.

LP Linear program. An optimization formulation where a problem is defined by a linear objective function, a set of linear equalities and a set of linear inequalities. FBA is a kind of LP.

MILP A LP with integer variables. The problem is then piecewise-linear, and requires specific solving methods. When all the integer variables are fixed, a LP is obtained. TFA is a kind of MILP.

Special Ordered Set of type 1 (SOS1) constraint A type of constraint where a sum of binary variables has to be lower than or equal to 1. Useful to model a choice between different possibilities.

Zeroth order approximation Approximation of a function using a piecewise constant function. The values of the zeroth-order approximation of the function are a discretization of the space of values of the initial function.

Supplementary References

- [1] 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 [Journal Article]. *Bioinformatics*. 2018;.
- [2] Dalke A, Wilczynski B, Chapman BA, Cox CJ, Kauff F, Friedberg I, et al. Biopython: freely available Python tools for computational molecular biology and bioinformatics. *Bioinformatics*. 2009 03;25(11):1422–1423. Available from: <https://dx.doi.org/10.1093/bioinformatics/btp163>.
- [3] Caspi R, Foerster H, Fulcher CA, Kaipa P, Krummenacker M, Latendresse M, et al. The MetaCyc Database of metabolic pathways and enzymes and the BioCyc collection of Pathway/Genome Databases. *Nucleic acids research*. 2007;36(suppl_1):D623–D631.
- [4] Keseler IM, Collado-Vides J, Gama-Castro S, Ingraham J, Paley S, Paulsen IT, et al. EcoCyc: a comprehensive database resource for *Escherichia coli*. *Nucleic acids research*. 2005;33(suppl_1):D334–D337.
- [5] Travers M, Paley SM, Shrager J, Holland TA, Karp PD. Groups: knowledge spreadsheets for symbolic biocomputing. *Database*. 2013;2013.
- [6] Fredrickson A. Formulation of structured growth models. *Biotechnology and bioengineering*. 1976;18(10):1481–1486.
- [7] 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.
- [8] Bartholomäus A, Fedyunin I, Feist P, Sin C, Zhang G, Valleriani A, et al. Bacteria differently regulate mRNA abundance to specifically respond to various stresses. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*. 2016;374(2063):20150069.
- [9] Taniguchi Y, Choi PJ, Li GW, Chen H, Babu M, Hearn J, et al. Quantifying *E. coli* proteome and transcriptome with single-molecule sensitivity in single cells. *Science*. 2010;329(5991):533–538.
- [10] Piques M, Schulze WX, Höhne M, Usadel B, Gibon Y, Rohwer J, et al. Ribosome and transcript copy numbers, polysome occupancy and enzyme dynamics in *Arabidopsis*. *Molecular systems biology*. 2009;5(1).
- [11] Bremer H, Dennis PP. Modulation of chemical composition and other parameters of the cell by growth rate [Journal Article]. *Escherichia coli and Salmonella: cellular and molecular biology*. 1996;2(2):1553–69.
- [12] Dill KA, Ghosh K, Schmit JD. Physical limits of cells and proteomes. *Proceedings of the National Academy of Sciences*. 2011;108(44):17876–17882.
- [13] Kaleta C, Schäuble S, Rinas U, Schuster S. Metabolic costs of amino acid and protein production in *Escherichia coli*. *Biotechnology journal*. 2013;8(9):1105–1114.

- [14] Stephanopoulos G, Aristidou AA, Nielsen J. Metabolic engineering: principles and methodologies. Elsevier; 1998.
- [15] Benyamini T, Folger O, Ruppin E, Shlomi T. Flux balance analysis accounting for metabolite dilution. *Genome biology*. 2010;11(4):R43.
- [16] Heineken F, Tsuchiya H, Aris R. On the mathematical status of the pseudo-steady state hypothesis of biochemical kinetics. *Mathematical Biosciences*. 1967;1(1):95–113.
- [17] Monod J. The growth of bacterial cultures. *Annual review of microbiology*. 1949;3(1):371–394.
- [18] Campbell A. Synchronization of cell division. *Bacteriological reviews*. 1957;21(4):263.
- [19] Hinshelwood CN. The chemical kinetics of the bacterial cell.; 1946.



B. dETFL: Supplementary information

Supplementary Information

Emergence of diauxie as an optimal growth strategy
under resource allocation constraints in cellular
metabolism

Salvy et al.

Supplementary note S1: Discussion on the assumptions in DynamicME

Lloyd *et al.* [1] developed an efficient ME-model for *E. coli*, and Yang *et al.* used it to formulate a dynamic analysis framework (dynamic-ME) [2] similar to dynamic flux balance analysis (dFBA) [3]. However, in order to handle the computational complexity of their model, they introduced a number of assumptions which resulted in several limitations to their model. In particular, the following items might be limiting:

- The standard solving procedure uses a dedicated quad-precision solver [4] and its assorted solving algorithm [5].
- Moreover, an important assumption in this method is that the dynamic algorithm approximates uptake fluxes bounds to be constant as long as the substrate in question is not depleted, which neglects the impact of kinetic laws at different substrate concentrations.
- It also assumes the proteome does not change during that time.
- Additionally, the efficiency of their formulation comes from the use of equality constraints between the metabolic fluxes and the catalytic availability of the enzymes.
- As a result, these models cannot predict the presence of enzymes that do not carry flux, as can be the case in a transition phase between two phenotypes.
- More importantly, this method does not tackle the problem of alternative solutions at a given time-step, and hence does not acknowledge the possibility of different time traces depending on the solution choice.
- Finally, this method does not allow modeling thermodynamics constraints.

Supplementary Table S2

Reaction	Reaction name	Enzyme Symbol	k_{cat} [s ⁻¹]
GLCabcpp	Glucose transport via the ABC system	GLCt2pp_ABC_18	120.0
GLCt2pp	Glucose transport via proton symport	GLCt2pp_GALP	40.1
GLCptspp	Glucose transport via PEP to Pyruvate PTS	GLCptspp_157	134.5
		GLCptspp_164	139.2
		GLCptspp_165	135.3
HEX1	Hexokinase (glucose:ATP)	HEX1_GLUCOKIN	279
LACZpp	β -galactosidase (periplasmic)	LACZpp_EG12013	58.1
LACZ	β -galactosidase (cytoplasmic)	LACZ_BETAGALACTOSID	211
LCTStpp	Lactose transport via proton symport	LCTStpp_LACY	37.5
		LCTSt3ipp_YDEA	35.0
		LCTSt3ipp_B0070	35.1
		LCTSt3ipp_B2170	35.2
GALKr	Galactokinase	GALKr_G7096	38.0
		GALKr_GALACTOKIN	34.3
UGLT	UDPgucose–hexose-1-phosphate uridylyltransferase	UGLT_GALACTURIDYLYLTRANS	62.0
UDPG4E	UDPgucose 4-epimerase	UDPG4E_UDPGLUCEPIM	128
GALabcpp	Galactose transport via the ABC system	GALabcpp_ABC_18	120.0
		GALabcpp_ABC_46	110.0
GALt2pp	Galactose transport via proton symport	GALt2pp_GALP	40.1

Table S1: Properties of glucose and lactose transporting reactions and enzymes. Reaction names from the original iJO1366 model [6]. Enzyme symbols adapted from Biocyc [7]. k_{cat} values taken from Lloyd *et al.* [1].

Supplementary Figures S3-6: enzyme levels of pathways depending on the preculture conditions

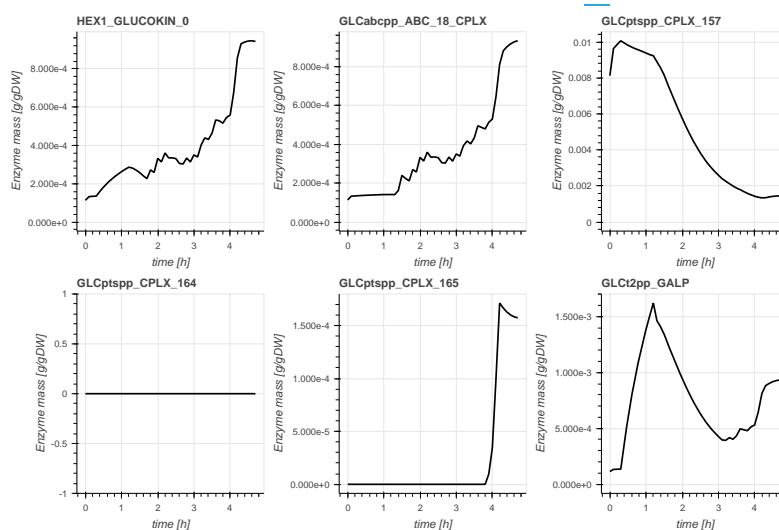


Figure S3: Enzyme levels of the glucose pathway, in the glucose/lactose diauxic experiment with glucose pre-culture.

Appendix B. dETFL: Supplementary information

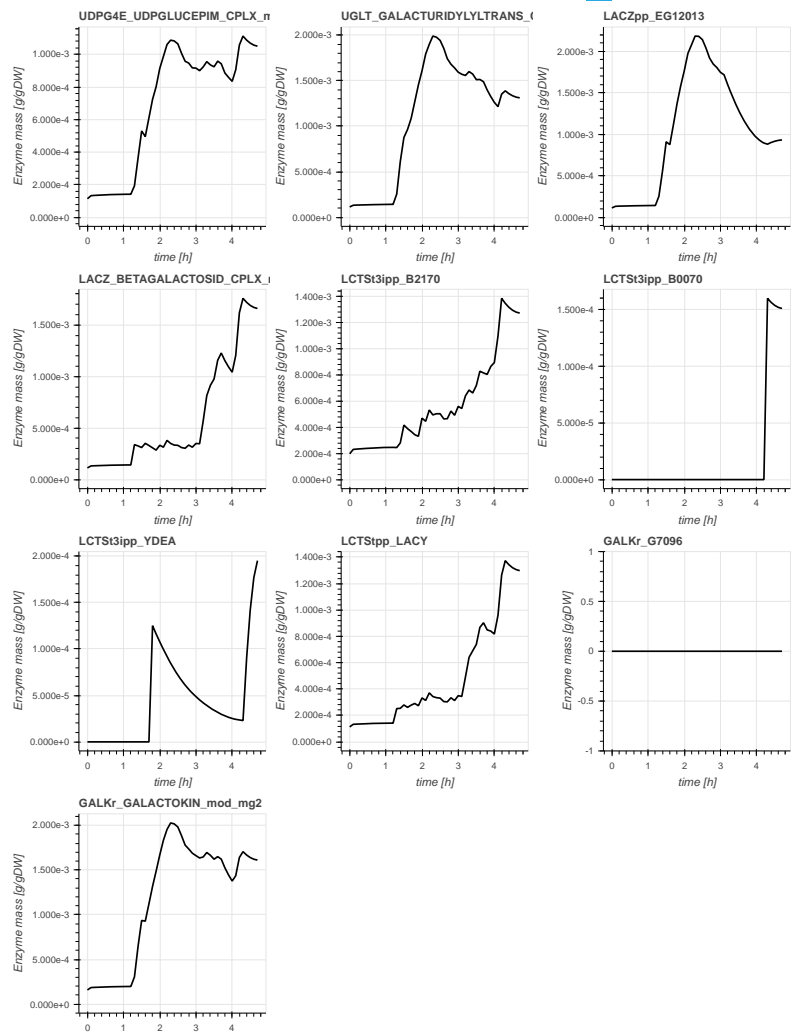


Figure S4: Enzyme levels of the lactose pathway, in the glucose/lactose diauxie experiment with glucose pre-culture.

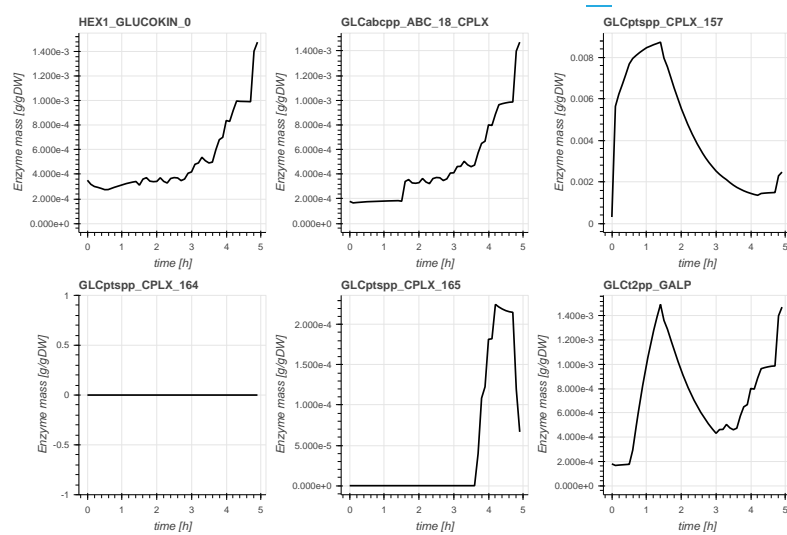


Figure S5: Enzyme levels of the glucose pathway, in the glucose/lactose diauxic experiment with lactose pre-culture.

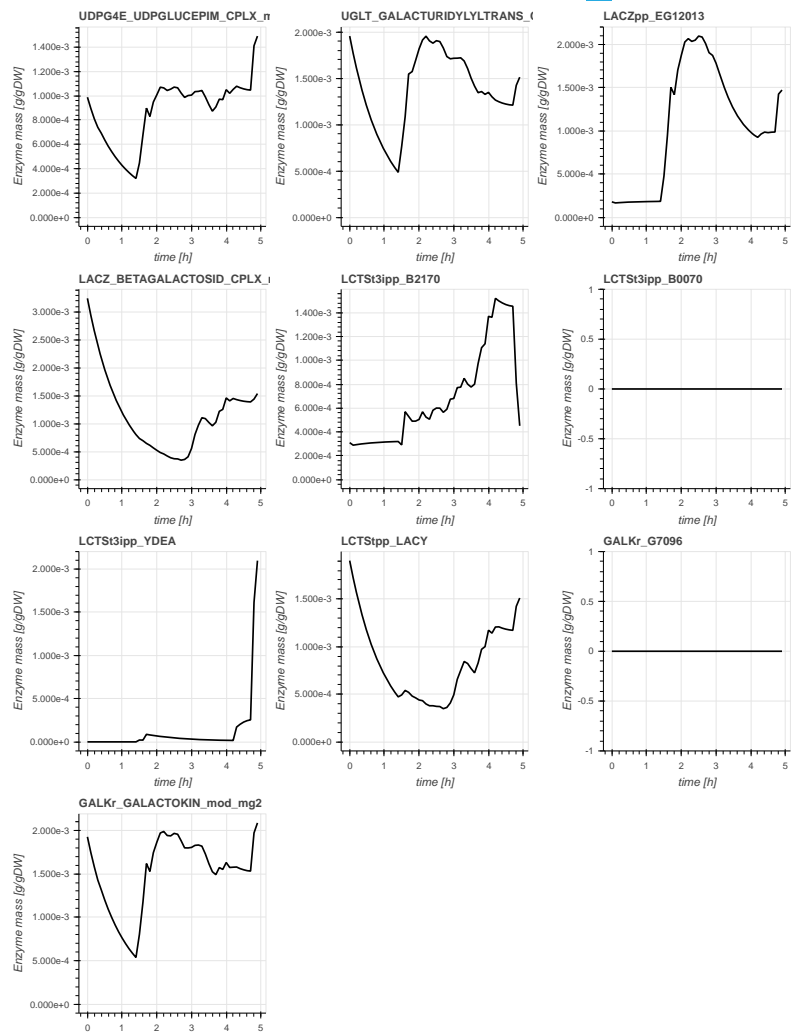
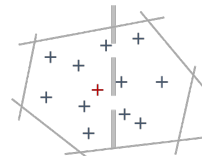


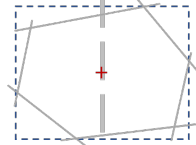
Figure S6: Enzyme levels of the lactose pathway, in the glucose/lactose diauxic experiment with lactose pre-culture.

Supplementary Figures S7-8: Chebyshev centering

a. Sampling + Mean



b. Variation Analysis + Mean



c. Chebyshev center:

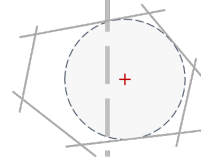


Figure S7: 2-Dimensional representation of different schemes to represent the solution space (gray polygon). The methods do not yield similar results (gray dashed line through the figures). **a.** It is possible to sample solutions (blue crosses) within the solution space, and take their mean (red cross) as a representative solution. The mean is still part of the solution space due to the convexity of the problem. **b.** Variation analysis (successive minimization/maximization of variables) allows to find the minimal bounding box (blue dashed lines) around the solution space. The center of the box (red cross) is also in the solution space, and can be used as a representative solution. **c.** The Chebyshev method finds the largest topological ball (grayed area) that can fit in the solution space. The center of the ball, or Chebyshev center (red cross) is also part of the solution space and can be used as a representative solution.

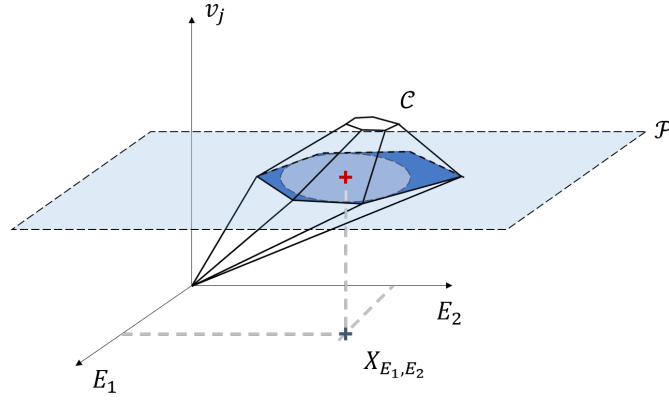


Figure S8: 3-Dimensional (E_1, E_2, v_j) example of a Chebyshev center. The feasible space is denoted by the polytope \mathcal{C} . The Chebyshev center with respect to variables E_1 and E_2 is X_{E_1, E_2} . It is the center of the largest 2-D sphere on a plane parallel to (E_1, E_2) that is inscribed in \mathcal{C} . This sphere exists on the plane \mathcal{P} , materialized in light blue.

Supplementary figure S9: Enzyme composition of the conceptual model when no constraints are applied to the rate-of-change of enzyme concentrations

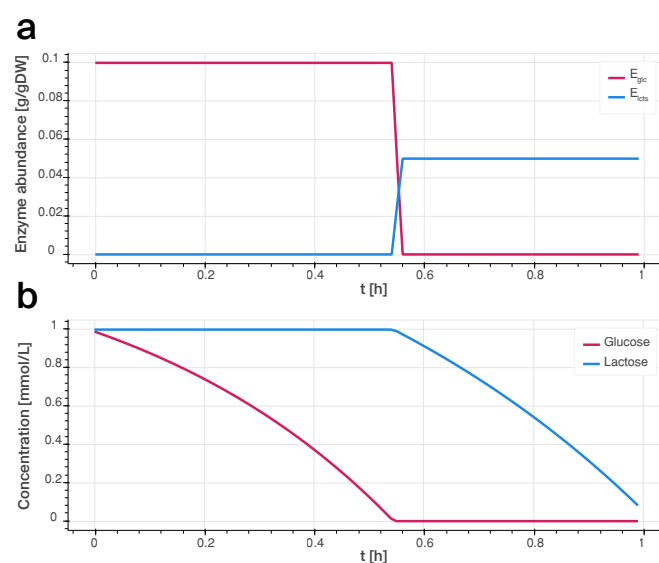


Figure S9: Enzyme composition of the conceptual model when no constraints are applied to the rate-of-change of enzyme concentrations. **a.** Enzyme content over time for the conceptual model on a mixed substrate. Glucose enzymes in pink, lactose enzymes in blue. **b.** Content of the batch reactor over time: Glucose (pink), lactose (blue).

Supplementary figure S10: dETFL results with switched k_{cat} between the glucose and Leloir pathways

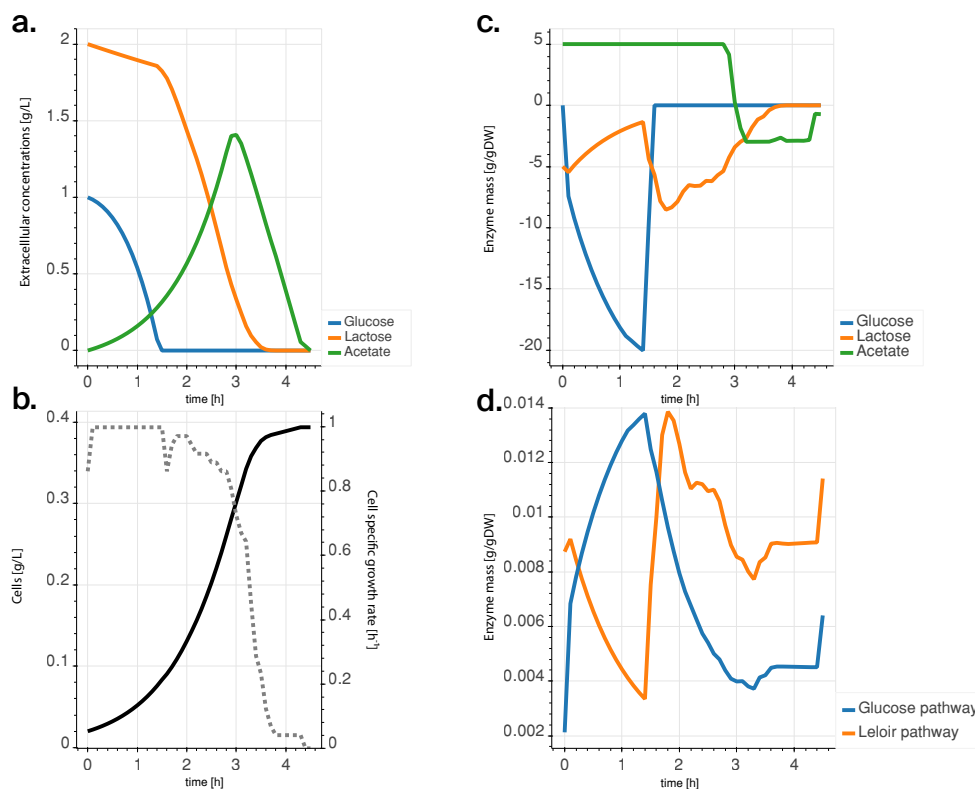


Figure S10: Results of the diauxic simulation with lactose-only preculture, and switched k_{cat} values between the glucose and Leloir pathways (resp. 37.7 s^{-1} and 135 s^{-1}): **a.** Temporal evolution of the extracellular concentrations of glucose (blue), lactose (orange), and acetate (green). **b.** Cell concentration (full line) and growth rate (dashed line) of the culture over time. **c.** Exchange rates of the cell, same colors as in subfigure -a. Positive exchange rates mean production, negative exchange rates mean consumption. **d.** Mass of enzymes allocated to the transformation of glucose (blue) and lactose (orange) in G6P. The dashed gray line shows the levels of β -galactosidase (LACZ) enzyme (in the Leloir pathway).

References

- [1] Lloyd CJ, Ebrahim A, Yang L, King ZA, Catoiu E, O'Brien EJ, et al. COBRAme: A computational framework for genome-scale models of metabolism and gene expression [Journal Article]. *PLoS computational biology*. 2018;14(7):e1006302.
- [2] Yang L, Ebrahim A, Lloyd CJ, Saunders MA, Palsson BO. DynamicME: dynamic simulation and refinement of integrated models of metabolism and protein expression. *BMC Systems Biology*. 2019 Jan;13(1):2. Available from: <https://doi.org/10.1186/s12918-018-0675-6>.
- [3] Mahadevan R, Edwards JS, Doyle r F J. Dynamic flux balance analysis of diauxic growth in *Escherichia coli* [Journal Article]. *Biophys J*. 2002;83(3):1331–40. Available from: <https://www.ncbi.nlm.nih.gov/pubmed/12202358>.
- [4] Ma D, Yang L, Fleming RM, Thiele I, Palsson BO, Saunders MA. Reliable and efficient solution of genome-scale models of Metabolism and macromolecular Expression. *Scientific reports*. 2017;7:40863.
- [5] Yang L, Ma D, Ebrahim A, Lloyd CJ, Saunders MA, Palsson BO. solveME: fast and reliable solution of nonlinear ME models [Journal Article]. *BMC Bioinformatics*. 2016;17(1):391. Available from: <https://www.ncbi.nlm.nih.gov/pubmed/27659412>.
- [6] Orth JD, Conrad TM, Na J, Lerman JA, Nam H, Feist AM, et al. A comprehensive genome-scale reconstruction of *Escherichia coli* metabolism-2011 [Journal Article]. *Molecular Systems Biology*. 2011;7. Available from: <GotoISI>://WQS:000296652600001.
- [7] Caspi R, Foerster H, Fulcher CA, Kaipa P, Krummenacker M, Latendresse M, et al. The MetaCyc Database of metabolic pathways and enzymes and the BioCyc collection of Pathway/Genome Databases. *Nucleic acids research*. 2007;36(suppl_1):D623–D631.



C. Regulation-enabled ETFL models of cancer: Supplementary information

Supplementary Information

Dose-dependent drug effect and resistance mechanisms
in a cancer model of metabolism and expression

Salvy et al.

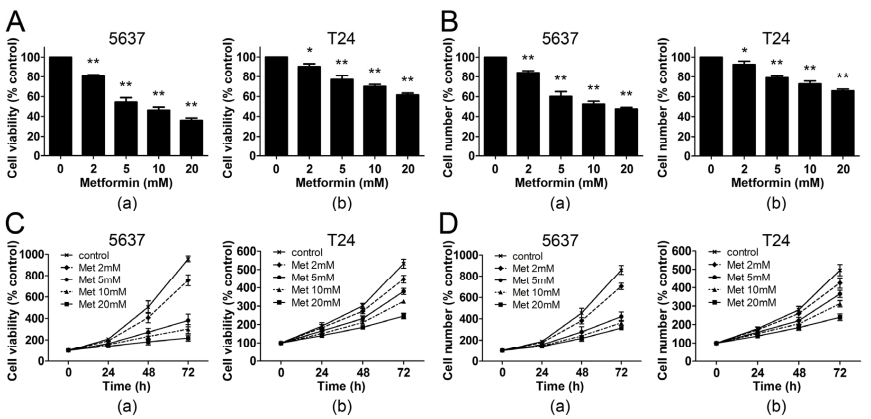
Supplementary Figure S1: In vitro cell viability after metformin treatment

Figure reprinted from:

Zhang T, Guo P, Zhang Y, Xiong H, Yu X, Xu S, et al. The antidiabetic drug metformin inhibits the proliferation of bladder cancer cells in vitro and in vivo. *International journal of molecular sciences*. 2013;14(12):24603–24618,

under the Creative Commons CC BY 4.0 license, as specified by the publisher (<https://www.mdpi.com/authors/rights>)

Figure 1. Metformin inhibits the proliferation of bladder cancer cells. (A) 5637 (a) and T24 (b) cells (5×10^3 cells/well) were seeded in 96-well culture plates. After 24 h, cells were treated with metformin (0, 2, 5, 10, 20 mM) for another 48 h. Cell viability was measured by MTT assay. The results were expressed as percent of cell viability compared with control (0 mM). Columns, means of three independent experiments; bars, SEs; (B) 5637 (a) and T24 (b) cells (5×10^4 cells/well) were seeded in 12-well culture plates. After treatment as in panel A, cell numbers were determined using a hemocytometer. The results were expressed as percent of viable cells compared with control. Columns, means of three independent experiments; bars, SEs; (C,D) 5637 (a) and T24 (b) cells were treated with metformin (Met) at different concentrations for 24, 48 and 72 h. Cell proliferation was measured by MTT (C) or cell count assay (D). Data, means of three independent experiments; bars, SEs. * $p < 0.05$ versus control; ** $p < 0.01$ versus control.





D. pyTFA & matTFA: Supplementary information

pyTFA & MatTFA: Supporting Note

1.1 Problem formulation

Thermodynamics Flux Analysis (TFA) adds constraints on top of a classic Flux Balance Analysis (FBA) problem to couple reaction directionalities to thermodynamics constraints. In particular, the formulation in (Soh and Hatzimanikatis, 2014) adds metabolite concentrations and Gibbs energy of reactions, and couples the sign of the Gibbs energy of a reaction to its directionality.

These constraints aim to reduce the feasible flux solution space of the problem and increasing the predictive power of the model. This methodology is used to perform Thermodynamics-based Variability Analysis (TVA), a series of TFA maximization and minimization of the variables in the model, such as reaction fluxes, to determine their allowable ranges and directionalities.

Given a model with specified reaction directionalities, it is possible to characterize the thermodynamic states of the underlying physiology by sampling equilibrium displacements and concentrations. We show here the formulation as proposed in (Soh and Hatzimanikatis, 2014):

FBA constraints	Mass balance	$S \cdot v = 0$
	Flux capacity	$v \leq v \leq \bar{v}$
TFA constraints	Gibbs energy of reaction	$\Delta_r G'_i = \Delta_{r,tpt} G'_i + \sum_{j=1}^m n_{i,j} \mu_j$
	Chemical potential	$\mu_j = \Delta_f G_j'^0 + \Delta_{f,err} G_j'^0 + RT \ln x_j$
	Thermodynamic feasibility	$\Delta_r G'_i - K + K * z_i < 0$
	Coupling constraint	$v_i - K * z_i < 0$

The TFA problem in the table incorporates thermodynamics-based constraints in the original FBA problem in the two first equations.

For biochemical reactions, the transformed Gibbs free energy of the reaction i , $\Delta_r G'_i$, is a function of the transformed Gibbs energy of the chemical potentials μ_j of the reactants j . If the reaction is a transport of the compounds from one compartment to another, the Gibbs free energy of transport $\Delta_{r,tpt} G'_i$ is also considered, according to the formulation in Jol *et al.* (Jol, et al., 2010). $\Delta_r G'_i$ is calculated in the third equation.

The chemical potential of the reactants is a function of the standard transformed Gibbs free energy of formation of the compounds $\Delta_f G_j'^0$ and the metabolite's activity, as shown in the fourth equation. Activities of the compounds can be expressed directly as concentrations, as we perform Debye-Hückel correction (Debye and Hückel, 1923). $\Delta_{f,err} G_j'^0$ is the estimated error in the energy of formation.

K is a large (Big-M, $K > \max \Delta_r G'_i$) value, and z is a binary variable. The two last equations enforce the constraint $\Delta_r G'_i < 0 \Leftrightarrow v_j \geq 0$. K should be chosen so that it is bigger by one or two orders of magnitude than the maximal abs ($\Delta_r G'_i$).

This formulation requires net fluxes to be non-negative. To do so, each reaction is separated in two: a net forward and a net backward, and their net fluxes are associated in the following manner:

$$v_{net} = v_{forward} - v_{backward}$$

In that form, the net forward and the net backward reactions are constrained to have non-negative values. Additional constraints are applied to ensure that at most one of these two is active at a time.

1.2 Usage and Example: sampling thermodynamic displacements

We provide a reduced *E. coli* model made with the software presented in (Ataman, et al., 2017), as well as the model it was generated from, iO1366 (Orth, et al., 2011).

We can sample the natural logarithms of thermodynamic displacement $\ln(\Gamma)$ for each reaction in the genome-scale model.

In a reaction with one product P and one substrate S, the thermodynamic displacement can be defined as such:

$$\Gamma = \frac{S_{eq} P}{P_{eq} S} = \frac{1}{k_{eq}} \frac{P}{S}$$

S_{eq}, P_{eq} are the concentration at equilibrium of the substrate and product according to the notations in (Heinrich and Schuster, 2012). k_{eq} is the associated equilibrium constant. In that context, we can also write the Gibbs energy of the reaction:

$$\begin{aligned} \frac{\Delta G}{RT} \\ \Delta G = \Delta G^0 + RT \ln\left(\frac{P}{S}\right) \\ \frac{\Delta G}{RT} = \frac{\Delta G^0}{RT} + \ln\left(\frac{P}{S}\right) \end{aligned}$$

Since $\ln\left(\frac{1}{k_{eq}}\right) = \frac{\Delta G^0}{RT}$:

$$e^{\frac{\Delta G}{RT}} = \frac{1}{k_{eq}} \frac{P}{S}$$

Hence,

$$\Gamma = e^{\frac{\Delta G}{RT}}$$

Then:

$$\Gamma = e^{\frac{\Delta G}{RT}} \Leftrightarrow \ln(\Gamma) = \frac{\Delta G}{RT}$$

and hence the directionality of a reaction is opposite to the sign of $\ln(\Gamma)$.

It is possible to directly sample admissible thermodynamics displacements and calculated $\overline{\ln \Gamma}$ the average thermodynamic displacement for each reaction. The thermodynamic displacements are constrained because of their link to $\Delta_r G^0$, which is defined by metabolite concentrations. Thus, admissible displacements depend directly on the concentration ranges.

It is also useful to sample directly admissible concentrations. As an example, here is a sampling performed for cytosolic ATP using the methods provided with the package:

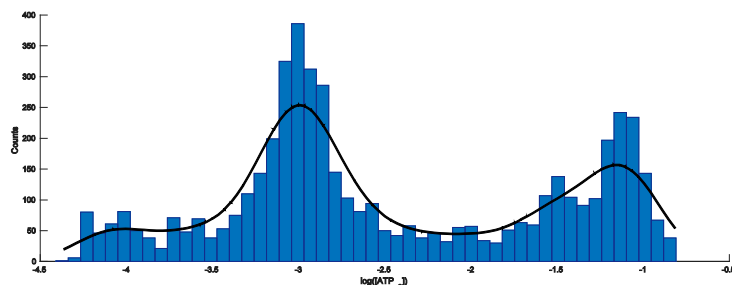


Figure S 1: Example distribution of the sampling of admissible ATP concentrations

1.3 Further Analysis

Given an MILP model with thermodynamics constraints, it is possible to perform several kinds of additional studies.

Thermodynamics-based Variability Analysis (TVA)

TVA can be performed on any variable of the model. These include metabolite concentrations, reaction fluxes, Gibbs free energy of reactions, displacement from equilibrium.

Integration of metabolomics

Integration of metabolomics data is possible because the logarithmic concentration of metabolites is a variable within the model. In particular, it is possible to perform Thermodynamics-based Metabolite Sensitivity Analysis (TMSA) (Kiparissides and Hatzimanikatis, 2017), which allows to define a priority list of metabolites to measure in order to constrain further the model.

Characterization of physiologies

By enumerating bidirectional reactions, and looking at the different solutions spanned by their directionalities, it is possible to characterize the relationship between different flux and physiologies. Additionally, it is possible to observe which reactions are operating close to or far from equilibrium.

Sampling

The resulting constraint-based model is amenable to sampling of any of its variables, such as metabolite concentrations or thermodynamic displacements. pyTFA and matTFA can indeed call COBRA's sampling methods, Artificially-Centered Hit and Run (Schellenberger and Palsson, 2009) and OptGpSampler (Megchelenbrink, et al., 2014). Given a physiology, this allows preparing data for kinetic modeling methods, such as Metabolic Control Analysis-based ORACLE (Miskovic, et al., 2017; Miskovic and Hatzimanikatis, 2010).

1.4 Data

Gibbs free energies of formation ΔG_f^0

Gibbs free energies of formation can be obtained from various data sources, among them:

- Literature, e.g. (Jankowski, et al., 2008)
- eQuilibrator (Flamholz, et al., 2012)
- Databases (eg NIST)
- LCSB also provides support on obtaining these data upon request.

1.5 References

- Ataman, M., *et al.* redGEM: Systematic reduction and analysis of genome-scale metabolic reconstructions for development of consistent core metabolic models. *Plos Comput Biol* 2017;13(7).
- Debye, P. and Hückel, E. Zur Theorie der Elektrolyte. I. Gefrierpunktserniedrigung und verwandte Erscheinungen. The theory of electrolytes. I. Lowering of freezing point and related phenomena. *Phys Z* 1923;24:185-206.
- Flamholz, A., *et al.* eQuilibrator--the biochemical thermodynamics calculator. *Nucleic Acids Res* 2012;40(Database issue):D770-775.
- Heinrich, R. and Schuster, S. The regulation of cellular systems. Springer Science & Business Media; 2012.
- Jankowski, M.D., *et al.* Group contribution method for thermodynamic analysis of complex metabolic networks. *Biophys J* 2008;95(3):1487-1499.
- Jol, S.J., *et al.* Thermodynamic Calculations for Biochemical Transport and Reaction Processes in Metabolic Networks. *Biophys J* 2010;99(10):3139-3144.
- Kiparissides, A. and Hatzimanikatis, V. Thermodynamics-based Metabolite Sensitivity Analysis in metabolic networks. *Metab Eng* 2017;39:117-127.
- Megchelenbrink, W., Huynen, M. and Marchiori, E. optGpSampler: an improved tool for uniformly sampling the solution-space of genome-scale metabolic networks. *Plos One* 2014;9(2):e86587.
- Miskovic, L., *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.
- Miskovic, L. and Hatzimanikatis, V. Production of biofuels and biochemicals: in need of an ORACLE. *Trends Biotechnol* 2010;28(8):391-397.
- Orth, J.D., *et al.* A comprehensive genome-scale reconstruction of *Escherichia coli* metabolism-2011. *Mol Syst Biol* 2011;7.
- Schellenberger, J. and Palsson, B.O. Use of randomized sampling for analysis of metabolic networks. *J Biol Chem* 2009;284(9):5457-5461.
- Soh, K.C. and Hatzimanikatis, V. Constraining the flux space using thermodynamics and integration of metabolomics data. *Methods Mol Biol* 2014;1191:49-63.



E. SKiMPy: Supplementary information

Supplementary Information

Symbolic kinetic models in Python: SKiMPy

Weilandt et al.

Reaction	Mechanism	Explicit rate law formulation
$A \rightleftharpoons B$	Reversible Michaelis-Menten	$v_r = V_{max,r} \frac{\frac{[A]_r}{K_{A,r}} (1 - \frac{[B]_r}{K_{eq,r}}) \frac{[B]_r}{[A]_r}}{1 + \frac{[A]_r}{K_{A,r}} + \frac{[B]_r}{K_{B,r}}}$
$A \rightleftharpoons B + C$	Uni - Bi Reversible Hill	$v_r = \frac{V_{max,r} \frac{[A]_r}{K_{A,r}} (1 - \frac{[B]_r}{K_{eq,r}}) \frac{[C]_r^{h_r}}{[A]_r} (\frac{[A]_r}{K_{A,r}} + \frac{[B]_r}{K_{B,r}} + \frac{[C]_r}{K_{C,r}})^{h_r-1} (\frac{[B]_r}{K_{B,r}} + \frac{[C]_r}{K_{C,r}})^{h_r-1}}{1 + (\frac{[A]_r}{K_{A,r}} + \frac{[B]_r}{K_{B,r}})^{h_r} + (\frac{[A]_r}{K_{A,r}} + \frac{[B]_r}{K_{B,r}} + \frac{[C]_r}{K_{C,r}})^{h_r} - 2(\frac{[A]_r}{K_{A,r}})^{h_r}}$
$A \rightleftharpoons B + C + D$	Convenience	$v_r = \frac{V_{max,r} \frac{[A]_r}{K_{A,r}} \frac{[B]_r}{K_{B,r}} (1 - \frac{[C]_r}{K_{eq,r}}) \frac{[D]_r}{[A]_r}}{(1 + \frac{[A]_r}{K_{A,r}}) + (1 + \frac{[B]_r}{K_{B,r}}) (1 + \frac{[C]_r}{K_{C,r}}) (1 + \frac{[D]_r}{K_{D,r}}) - 1}$
$A + B \rightleftharpoons C + D$	Generalized Reversible Hill	$v_r = \frac{V_{max,r} \frac{[A]_r}{K_{A,r}} \frac{[B]_r}{K_{B,r}} (1 - \frac{[C]_r}{K_{eq,r}}) \frac{[D]_r^{h_r}}{[A]_r [B]_r} (\frac{[A]_r}{K_{A,r}} + \frac{[B]_r}{K_{B,r}} + \frac{[C]_r}{K_{C,r}})^{h_r-1} (\frac{[B]_r}{K_{B,r}} + \frac{[D]_r}{K_{D,r}})^{h_r-1}}{(1 + (\frac{[A]_r}{K_{A,r}} + \frac{[C]_r}{K_{C,r}})^{h_r}) (1 + (\frac{[B]_r}{K_{B,r}} + \frac{[D]_r}{K_{D,r}})^{h_r})}$
$A + B \rightleftharpoons C + D + E$	Convenience	$v_r = \frac{V_{max,r} \frac{[A]_r}{K_{A,r}} \frac{[B]_r}{K_{B,r}} (1 - \frac{[C]_r}{K_{eq,r}}) \frac{[D]_r}{[A]_r} \frac{[E]_r}{[B]_r}}{(1 + \frac{[A]_r}{K_{A,r}}) (1 + \frac{[B]_r}{K_{B,r}}) + (1 + \frac{[C]_r}{K_{C,r}}) (1 + \frac{[D]_r}{K_{D,r}}) (1 + \frac{[E]_r}{K_{E,r}}) - 1}$
$A + B \rightleftharpoons C + D + E + F$	Convenience	$v_r = \frac{V_{max,r} \frac{[A]_r}{K_{A,r}} \frac{[B]_r}{K_{B,r}} (1 - \frac{[C]_r}{K_{eq,r}}) \frac{[D]_r}{[A]_r} \frac{[E]_r}{[B]_r} \frac{[F]_r}{[C]_r}}{(1 + \frac{[A]_r}{K_{A,r}}) (1 + \frac{[B]_r}{K_{B,r}}) + (1 + \frac{[C]_r}{K_{C,r}}) (1 + \frac{[D]_r}{K_{D,r}}) (1 + \frac{[E]_r}{K_{E,r}}) (1 + \frac{[F]_r}{K_{F,r}}) - 1}$
$A + B \rightleftharpoons 2C + D + E$	Convenience	$v_r = \frac{V_{max,r} \frac{[A]_r}{K_{A,r}} \frac{[B]_r}{K_{B,r}} (1 - \frac{[C]_r}{K_{eq,r}}) \frac{[D]_r}{[A]_r} \frac{[E]_r}{[B]_r}}{(1 + \frac{[A]_r}{K_{A,r}}) (1 + \frac{[B]_r}{K_{B,r}}) + (1 + \frac{[C]_r}{K_{C,r}}) (\frac{[C]_r}{K_{C,r}})^2 (1 + \frac{[D]_r}{K_{D,r}}) (1 + \frac{[E]_r}{K_{E,r}}) - 1}$
$A + B \rightleftharpoons 3C + D + E$	Convenience	$v_r = \frac{V_{max,r} \frac{[A]_r}{K_{A,r}} \frac{[B]_r}{K_{B,r}} (1 - \frac{[C]_r}{K_{eq,r}}) \frac{[D]_r}{[A]_r} \frac{[E]_r}{[B]_r}}{(1 + \frac{[A]_r}{K_{A,r}}) (1 + \frac{[B]_r}{K_{B,r}}) + (1 + \frac{[C]_r}{K_{C,r}}) (\frac{[C]_r}{K_{C,r}})^3 (1 + \frac{[D]_r}{K_{D,r}}) (1 + \frac{[E]_r}{K_{E,r}}) - 1}$

Table S1: Types of reaction, mechanisms and explicit rate law formulations used to compute kinetic parameters sets — Part 1, from Weilandt & Masid, Bernard-Bruels, *Kinetic models to study the Warburg effect in cancer cells*, Unpublished manuscript.

Reaction	Mechanism	Explicit rate law formulation
$A + B + C \rightleftharpoons D$	Convenience	$v_r = \frac{V_{max,r} \frac{[A]_r [B]_r [C]_r}{K_{A,r} K_{B,r} K_{C,r}} (1 - \frac{[D]_r}{K_{eq,r} [A]_r [B]_r [C]_r})}{(1 + \frac{[A]_r}{K_{A,r}}) (1 + \frac{[B]_r}{K_{B,r}}) (\frac{[C]_r}{K_{C,r}}) + (1 + \frac{[D]_r}{K_{D,r}}) - 1}$
$A + B + C \rightleftharpoons D + E$	Convenience	$v_r = \frac{V_{max,r} \frac{[A]_r [B]_r [C]_r}{K_{A,r} K_{B,r} K_{C,r}} (1 - \frac{[D]_r [E]_r}{K_{eq,r} [A]_r [B]_r [C]_r})}{(1 + \frac{[A]_r}{K_{A,r}}) (1 + \frac{[B]_r}{K_{B,r}}) (\frac{[C]_r}{K_{C,r}}) + (1 + \frac{[D]_r}{K_{D,r}}) (1 + \frac{[E]_r}{K_{E,r}}) - 1}$
$A + B + C \rightleftharpoons D + E + F$	Generalized Reversible Hill	$v_r = \frac{V_{max,r} \frac{[A]_r [B]_r [C]_r}{K_{A,r} K_{B,r} K_{C,r}} (1 - \frac{[D]_r [E]_r [F]_r}{K_{eq,r} [A]_r [B]_r [C]_r})^{h_r-1} (\frac{[D]_r}{K_{D,r}} + \frac{[E]_r}{K_{E,r}} + \frac{[F]_r}{K_{F,r}})^{h_r-1}}{(1 + \frac{[A]_r}{K_{A,r}} + \frac{[D]_r}{K_{D,r}})^{h_r} (1 + \frac{[B]_r}{K_{B,r}} + \frac{[E]_r}{K_{E,r}})^{h_r} (1 + \frac{[C]_r}{K_{C,r}} + \frac{[F]_r}{K_{F,r}})^{h_r}}$
$A + B + C \rightleftharpoons 2D + E + F$	Convenience	$v_r = \frac{V_{max,r} \frac{[A]_r [B]_r [C]_r}{K_{A,r} K_{B,r} K_{C,r}} (1 - \frac{[D]_r^2 [E]_r [F]_r}{K_{eq,r} [A]_r [B]_r [C]_r})}{(1 + \frac{[A]_r}{K_{A,r}}) (1 + \frac{[B]_r}{K_{B,r}}) (\frac{[C]_r}{K_{C,r}}) + (1 + \frac{[D]_r}{K_{D,r}})^2 (1 + \frac{[E]_r}{K_{E,r}}) (1 + \frac{[F]_r}{K_{F,r}}) - 1}$
$2A + B + C \rightleftharpoons 2D + E$	Convenience	$v_r = \frac{V_{max,r} (\frac{[A]_r}{K_{A,r}})^2 \frac{[B]_r [C]_r}{K_{B,r} K_{C,r}} (1 - \frac{[D]_r^2 [E]_r}{K_{eq,r} [A]_r^2 [B]_r [C]_r})}{(1 + \frac{[A]_r}{K_{A,r}} + (\frac{[A]_r}{K_{A,r}})^2) (1 + \frac{[B]_r}{K_{B,r}}) (\frac{[C]_r}{K_{C,r}}) + (1 + \frac{[D]_r}{K_{D,r}})^2 (1 + \frac{[E]_r}{K_{E,r}}) - 1}$
$3A + B \rightleftharpoons 3C + D$	Generalized Reversible Hill	$v_r = \frac{V_{max,r} (\frac{[A]_r}{K_{A,r}})^3 \frac{[B]_r}{K_{B,r}} (1 - \frac{[C]_r^3 [D]_r}{K_{eq,r} [A]_r^3 [B]_r})^{h_r-1} (\frac{[C]_r}{K_{C,r}} + \frac{[D]_r}{K_{D,r}})^{h_r-1}}{(1 + (\frac{[A]_r}{K_{A,r}} + \frac{[C]_r}{K_{C,r}})^{h_r})^3 (1 + \frac{[B]_r}{K_{B,r}} + \frac{[D]_r}{K_{D,r}})^{h_r}}$
$2A + B + C \rightleftharpoons 2D + E + F + G + H$	Convenience	$v_r = \frac{V_{max,r} (\frac{[A]_r}{K_{A,r}})^2 \frac{[B]_r [C]_r}{K_{B,r} K_{C,r}} (1 - \frac{[D]_r^2 [E]_r [F]_r [G]_r [H]_r}{K_{eq,r} [A]_r^2 [B]_r [C]_r})}{(1 + \frac{[A]_r}{K_{A,r}} + (\frac{[A]_r}{K_{A,r}})^2) (1 + \frac{[B]_r}{K_{B,r}}) (\frac{[C]_r}{K_{C,r}}) + (1 + \frac{[D]_r}{K_{D,r}} + \frac{[E]_r}{K_{E,r}})^2 (1 + \frac{[F]_r}{K_{F,r}}) (1 + \frac{[G]_r}{K_{G,r}}) (1 + \frac{[H]_r}{K_{H,r}}) - 1}$
$A + B + C + D + E \rightleftharpoons F + G + H$	Convenience	$v_r = \frac{V_{max,r} (\frac{[A]_r [B]_r [C]_r [D]_r [E]_r}{K_{A,r} K_{B,r} K_{C,r} K_{D,r} K_{E,r}} (1 - \frac{[F]_r [G]_r [H]_r}{K_{eq,r} [A]_r [B]_r [C]_r [D]_r [E]_r})}{(1 + \frac{[A]_r}{K_{A,r}}) (1 + \frac{[B]_r}{K_{B,r}}) (\frac{[C]_r}{K_{C,r}}) (1 + \frac{[D]_r}{K_{D,r}}) (1 + \frac{[E]_r}{K_{E,r}}) + (1 + \frac{[F]_r}{K_{F,r}}) (1 + \frac{[G]_r}{K_{G,r}}) (1 + \frac{[H]_r}{K_{H,r}}) - 1}$

Table S2: Types of reaction, mechanisms and explicit rate law formulations used to compute kinetic parameters sets - Part 2, from Weilandt & Masid, Bernard-Bruels, *Kinetic models to study the Warburg effect in cancer cells*, Unpublished manuscript.

Bibliography

- [1] J. E. Bailey, “Toward a science of metabolic engineering,” *Science*, vol. 252, no. 5013, pp. 1668–1675, 1991.
- [2] D. M. Pearsall, “Plant domestication and the shift to agriculture in the andes,” in *The handbook of South American archaeology*, pp. 105–120, Springer, 2008.
- [3] G. Mendel, A. F. Corcos, and F. V. Monaghan, *Gregor Mendel’s experiments on plant hybrids: a guided study*. Rutgers University Press, 1993.
- [4] W. Klümper and M. Qaim, “A meta-analysis of the impacts of genetically modified crops,” *PloS one*, vol. 9, no. 11, p. e111629, 2014.
- [5] K. E. Ormond, D. P. Mortlock, D. T. Scholes, Y. Bombard, L. C. Brody, W. A. Faucett, G. Nanibaa’A, L. Hercher, R. Isasi, A. Middleton, *et al.*, “Human germline genome editing,” *The American Journal of Human Genetics*, vol. 101, no. 2, pp. 167–176, 2017.
- [6] D. K. Ro, E. M. Paradise, M. Ouellet, K. J. Fisher, K. L. Newman, J. M. Ndungu, K. A. Ho, R. A. Eachus, T. S. Ham, J. Kirby, M. C. Y. Chang, S. T. Withers, Y. Shiba, R. Sarpong, and J. D. Keasling, “Production of the antimalarial drug precursor artemisinic acid in engineered yeast,” *Nature*, vol. 440, no. 7086, pp. 940–943, 2006.
- [7] A. L. Meadows, K. M. Hawkins, Y. Tsegaye, E. Antipov, Y. Kim, L. Raetz, R. H. Dahl, A. Tai, T. Mahatdejkul-Meadows, L. Xu, L. S. Zhao, M. S. Dasika, A. Murarka, J. Lenihan, D. Eng, J. S. Leng, C. L. Liu, J. W. Wenger, H. X. Jiang, L. L. Chao, P. Westfall, J. Lai, S. Ganesan, P. Jackson, R. Mans, D. Platt, C. D. Reeves, P. R. Saija, G. Wichmann, V. F. Holmes, K. Benjamin, P. W. Hill, T. S. Gardner, and A. E. Tsong, “Rewriting yeast central carbon metabolism for industrial isoprenoid production,” *Nature*, vol. 537, no. 7622, pp. 694–+, 2016.
- [8] D. Mendez-Perez, J. Alonso-Gutierrez, Q. Hu, M. Molinas, E. E. Baidoo, G. Wang, L. J. Chan, P. D. Adams, C. J. Petzold, J. D. Keasling, *et al.*, “Production of jet fuel precursor monoterpenoids from engineered *escherichia coli*,” *Biotechnology and bioengineering*, vol. 114, no. 8, pp. 1703–1712, 2017.

Bibliography

- [9] P. Xu, K. J. Qiao, W. S. Ahn, and G. Stephanopoulos, “Engineering *yarrowia lipolytica* as a platform for synthesis of drop-in transportation fuels and oleochemicals,” *Proceedings of the National Academy of Sciences of the United States of America*, vol. 113, no. 39, pp. 10848–10853, 2016.
- [10] M. Coelho, P. Amaral, and I. Belo, “*Yarrowia lipolytica*: an industrial workhorse,” *Current research, technology and education topics in applied microbiology and microbial biotechnology*, vol. 2, pp. 930–940, 2010.
- [11] F. A. Goncalves, G. Colen, and J. A. Takahashi, “*Yarrowia lipolytica* and its multiple applications in the biotechnological industry,” *ScientificWorldJournal*, vol. 2014, p. 476207, 2014.
- [12] N. Savić and G. Schwank, “Advances in therapeutic crispr/cas9 genome editing,” *Translational Research*, vol. 168, pp. 15–21, 2016.
- [13] T. K. MacLachlan, M. Lukason, M. Collins, R. Munger, E. Isenberger, C. Rogers, S. Malatos, E. DuFresne, J. Morris, R. Calcedo, *et al.*, “Preclinical safety evaluation of aav2-sft01—a gene therapy for age-related macular degeneration,” *Molecular Therapy*, vol. 19, no. 2, pp. 326–334, 2011.
- [14] E. P. Rakoczy, C.-M. Lai, A. L. Magno, M. E. Wikstrom, M. A. French, C. M. Pierce, S. D. Schwartz, M. S. Blumenkranz, T. W. Chalberg, M. A. Degli-Esposti, *et al.*, “Gene therapy with recombinant adeno-associated vectors for neovascular age-related macular degeneration: 1 year follow-up of a phase 1 randomised clinical trial,” *The Lancet*, vol. 386, no. 10011, pp. 2395–2403, 2015.
- [15] M. Gore, “Adverse effects of gene therapy: gene therapy can cause leukaemia: no shock, mild horror but a probe,” *Gene Therapy*, vol. 10, no. 1, p. 4, 2003.
- [16] M. R. Douglas and J. F. Tooker, “Large-scale deployment of seed treatments has driven rapid increase in use of neonicotinoid insecticides and preemptive pest management in us field crops,” *Environmental science & technology*, vol. 49, no. 8, pp. 5088–5097, 2015.
- [17] J. E. Bailey, “Complex biology with no parameters,” *Nature biotechnology*, vol. 19, no. 6, pp. 503–504, 2001.
- [18] M. MOLINAS, R. FAIZOVA, M. MAZZANTI, B. SCHACHERL, J. GALANZEW, T. VITOVA, and R. BERNIERLATMANI, “U (v) is an intermediate in the reduction of u (vi) by *shewanella oneidensis* mr-1,”
- [19] A. BROWN, M. MOLINAS, Y. ROEBBERT, R. FAIZOVA, S. W. MAZZANTI, and R. BERNIERLATMANI, “Uranium speciation impacts isotope signatures arising from microbial u (vi) reduction,”

-
- [20] F. R. Blattner, G. Plunkett, C. A. Bloch, N. T. Perna, V. Burland, M. Riley, J. Collado-Vides, J. D. Glasner, C. K. Rode, G. F. Mayhew, *et al.*, “The complete genome sequence of escherichia coli k-12,” *science*, vol. 277, no. 5331, pp. 1453–1462, 1997.
- [21] A. Goffeau, B. G. Barrell, H. Bussey, R. W. Davis, B. Dujon, H. Feldmann, F. Galibert, J. D. Hoheisel, C. Jacq, M. Johnston, E. J. Louis, H. W. Mewes, Y. Murakami, P. Philippsen, H. Tettelin, and S. G. Oliver, “Life with 6000 genes,” *Science*, vol. 274, no. 5287, pp. 546–567, 1996.
- [22] J. C. Venter, M. D. Adams, E. W. Myers, P. W. Li, R. J. Mural, G. G. Sutton, H. O. Smith, M. Yandell, C. A. Evans, R. A. Holt, *et al.*, “The sequence of the human genome,” *science*, vol. 291, no. 5507, pp. 1304–1351, 2001.
- [23] G. J. E. Baart and D. E. Martens, *Genome-Scale Metabolic Models: Reconstruction and Analysis*, pp. 107–126. Totowa, NJ: Humana Press, 2012.
- [24] C. Gu, G. B. Kim, W. J. Kim, H. U. Kim, and S. Y. Lee, “Current status and applications of genome-scale metabolic models,” *Genome biology*, vol. 20, no. 1, p. 121, 2019.
- [25] V. Chelliah, N. Juty, I. Ajmera, R. Ali, M. Dumousseau, M. Glont, M. Hucka, G. Jalowicki, S. Keating, V. Knight-Schrijver, A. Lloret-Villas, K. N. Natarajan, J. B. Pettit, N. Rodriguez, M. Schubert, S. M. Wimalaratne, Y. Y. Zhao, H. Hermjakob, N. Le Novere, and C. Laibe, “Biomodels: ten-year anniversary,” *Nucleic Acids Research*, vol. 43, no. D1, pp. D542–D548, 2015.
- [26] Z. A. King, J. Lu, A. Drager, P. Miller, S. Federowicz, J. A. Lerman, A. Ebrahim, B. O. Palsson, and N. E. Lewis, “Bigg models: A platform for integrating, standardizing and sharing genome-scale models,” *Nucleic Acids Research*, vol. 44, no. D1, pp. D515–D522, 2016.
- [27] A. Varma and B. O. Palsson, “Stoichiometric flux balance models quantitatively predict growth and metabolic by-product secretion in wild-type escherichia coli w3110,” *Appl. Environ. Microbiol.*, vol. 60, no. 10, pp. 3724–3731, 1994.
- [28] J. D. Orth, I. Thiele, and B. O. Palsson, “What is flux balance analysis?,” *Nature Biotechnology*, vol. 28, no. 3, pp. 245–248, 2010.
- [29] J. S. Edwards, R. U. Ibarra, and B. O. Palsson, “In silico predictions of escherichia coli metabolic capabilities are consistent with experimental data,” *Nature Biotechnology*, vol. 19, no. 2, pp. 125–130, 2001.
- [30] A. R. Zomorodi, P. F. Suthers, S. Ranganathan, and C. D. Maranas, “Mathematical optimization applications in metabolic networks,” *Metab Eng*, vol. 14, no. 6, pp. 672–86, 2012.

Bibliography

- [31] R. Schuetz, L. Kuepfer, and U. Sauer, “Systematic evaluation of objective functions for predicting intracellular fluxes in escherichia coli,” *Molecular Systems Biology*, vol. 3, 2007.
- [32] E. P. Gianchandani, M. A. Oberhardt, A. P. Burgard, C. D. Maranas, and J. A. Papin, “Predicting biological system objectives de novo from internal state measurements,” *Bmc Bioinformatics*, vol. 9, 2008.
- [33] N. E. Lewis, H. Nagarajan, and B. O. Palsson, “Constraining the metabolic genotype–phenotype relationship using a phylogeny of in silico methods,” *Nature Reviews Microbiology*, vol. 10, no. 4, pp. 291–305, 2012.
- [34] A. Chiappino-Pepe, V. Pandey, M. Ataman, and V. Hatzimanikatis, “Integration of metabolic, regulatory and signaling networks towards analysis of perturbation and dynamic responses,” *Current Opinion in Systems Biology*, vol. 2, pp. 59–66, 2017.
- [35] C. S. Henry, L. J. Broadbelt, and V. Hatzimanikatis, “Thermodynamics-based metabolic flux analysis,” *Biophysical Journal*, vol. 92, no. 5, pp. 1792–1805, 2007.
- [36] K. C. Soh and V. Hatzimanikatis, “Constraining the flux space using thermodynamics and integration of metabolomics data,” *Methods Mol Biol*, vol. 1191, pp. 49–63, 2014.
- [37] Q. K. Beg, A. Vazquez, J. Ernst, M. A. de Menezes, Z. Bar-Joseph, A. L. Barabasi, and Z. N. Oltvai, “Intracellular crowding defines the mode and sequence of substrate uptake by escherichia coli and constrains its metabolic activity,” *Proceedings of the National Academy of Sciences of the United States of America*, vol. 104, no. 31, pp. 12663–12668, 2007.
- [38] B. J. Sanchez, C. Zhang, A. Nilsson, P. J. Lahtvee, E. J. Kerkhoven, and J. Nielsen, “Improving the phenotype predictions of a yeast genome-scale metabolic model by incorporating enzymatic constraints,” *Molecular Systems Biology*, vol. 13, no. 8, 2017.
- [39] J. A. Lerman, D. R. Hyduke, H. Latif, V. A. Portnoy, N. E. Lewis, J. D. Orth, A. C. Schrimpe-Rutledge, R. D. Smith, J. N. Adkins, K. Zengler, *et al.*, “In silico method for modelling metabolism and gene product expression at genome scale,” *Nature communications*, vol. 3, p. 929, 2012.
- [40] E. J. O’Brien, J. A. Lerman, R. L. Chang, D. R. Hyduke, and B. O. Palsson, “Genome-scale models of metabolism and gene expression extend and refine growth phenotype prediction,” *Mol Syst Biol*, vol. 9, p. 693, 2013.
- [41] C. J. Lloyd, A. Ebrahim, L. Yang, Z. A. King, E. Catoiu, E. J. O’Brien, J. K. Liu, and B. O. Palsson, “Cobrame: A computational framework for genome-scale models of metabolism and gene expression,” *PLoS computational biology*, vol. 14, no. 7, p. e1006302, 2018.

-
- [42] A. J. Lopatkin and J. J. Collins, “Predictive biology: modelling, understanding and harnessing microbial complexity,” *Nature Reviews Microbiology*, pp. 1–14, 2020.
- [43] P. Salvy and V. Hatzimanikatis, “The etfl formulation allows multi-omics integration in thermodynamics-compliant metabolism and expression models,” *Nature Communications*, vol. 11, no. 1, pp. 1–17, 2020.
- [44] S. Magnúsdóttir, A. Heinken, L. Kutt, D. A. Ravcheev, E. Bauer, A. Noronha, K. Greenhalgh, C. Jäger, J. Baginska, P. Wilmes, *et al.*, “Generation of genome-scale metabolic reconstructions for 773 members of the human gut microbiota,” *Nature biotechnology*, vol. 35, no. 1, p. 81, 2017.
- [45] L. Yang, D. Ma, A. Ebrahim, C. J. Lloyd, M. A. Saunders, and B. O. Palsson, “solveME: fast and reliable solution of nonlinear me models,” *BMC Bioinformatics*, vol. 17, no. 1, p. 391, 2016.
- [46] D. Ma, L. Yang, R. M. Fleming, I. Thiele, B. O. Palsson, and M. A. Saunders, “Reliable and efficient solution of genome-scale models of metabolism and macromolecular expression,” *Scientific reports*, vol. 7, p. 40863, 2017.
- [47] F. C. Neidhardt and R. Curtiss, *Escherichia coli and Salmonella: cellular and molecular biology*, vol. 2. ASM press Washington, DC:, 1999.
- [48] J. D. Orth, T. M. Conrad, J. Na, J. A. Lerman, H. Nam, A. M. Feist, and B. O. Palsson, “A comprehensive genome-scale reconstruction of escherichia coli metabolism-2011,” *Molecular Systems Biology*, vol. 7, 2011.
- [49] D. McCloskey, J. A. Gangoiti, Z. A. King, R. K. Naviaux, B. A. Barshop, B. O. Palsson, and A. M. Feist, “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*, vol. 111, no. 4, pp. 803–815, 2014.
- [50] W. Liebermeister, E. Noor, A. Flamholz, D. Davidi, J. Bernhardt, and R. Milo, “Visual account of protein investment in cellular functions,” *Proceedings of the National Academy of Sciences of the United States of America*, vol. 111, no. 23, pp. 8488–8493, 2014.
- [51] A. Otto, J. Bernhardt, H. Meyer, M. Schaffer, F. A. Herbst, J. Siebourg, U. Mader, M. Lalk, M. Hecker, and D. Becher, “Systems-wide temporal proteomic profiling in glucose-starved bacillus subtilis,” *Nature Communications*, vol. 1, 2010.
- [52] J. Schellenberger and B. O. Palsson, “Use of randomized sampling for analysis of metabolic networks,” *J Biol Chem*, vol. 284, no. 9, pp. 5457–61, 2009.
- [53] J. Schellenberger, R. Que, R. M. T. Fleming, I. Thiele, J. D. Orth, A. M. Feist, D. C. Zielinski, A. Bordbar, N. E. Lewis, S. Rahmanian, J. Kang, D. R. Hyduke, and

Bibliography

- B. O. Palsson, "Quantitative prediction of cellular metabolism with constraint-based models: the cobra toolbox v2.0," *Nature Protocols*, vol. 6, no. 9, pp. 1290–1307, 2011.
- [54] W. Megchelenbrink, M. Huynen, and E. Marchiori, "optgpsampler: an improved tool for uniformly sampling the solution-space of genome-scale metabolic networks," *PLoS One*, vol. 9, no. 2, p. e86587, 2014.
- [55] J. Lee, W. Lam, and R. Dechter, "Benchmark on daoopt and gurobi with the pascal2 inference challenge problems," 2013.
- [56] A. Lodi and A. Tramontani, "Performance variability in mixed-integer programming," in *Theory Driven by Influential Applications*, pp. 1–12, INFORMS, 2013.
- [57] I. I. I. CPLEX, "High-performance mathematical programming engine," *International Business Machines Corp*, 2010.
- [58] Z. Gu, E. Rothberg, and R. Bixby, "Gurobi optimizer reference manual, version 8.0," *Gurobi Optimization Inc., Houston, USA*, 2018.
- [59] V. Pandey, N. Hadadi, and V. Hatzimanikatis, "Enhanced flux prediction by integrating relative expression and relative metabolite abundance into thermodynamically consistent metabolic models," *bioRxiv*, p. 481499, 2018.
- [60] H. Zur, E. Rupp, and T. Shlomi, "imat: an integrative metabolic analysis tool," *Bioinformatics*, vol. 26, no. 24, pp. 3140–3142, 2010.
- [61] S. A. Becker and B. O. Palsson, "Context-specific metabolic networks are consistent with experiments," *PLoS computational biology*, vol. 4, no. 5, p. e1000082, 2008.
- [62] V. Pandey and V. Hatzimanikatis, "Investigating the deregulation of metabolic tasks via minimum network enrichment analysis (minea) as applied to nonalcoholic fatty liver disease using mouse and human omics data," *bioRxiv*, p. 402222, 2018.
- [63] D. Segre, D. Vitkup, and G. M. Church, "Analysis of optimality in natural and perturbed metabolic networks," *Proceedings of the National Academy of Sciences*, vol. 99, no. 23, pp. 15112–15117, 2002.
- [64] R. Mahadevan, J. S. Edwards, and r. Doyle, F. J., "Dynamic flux balance analysis of diauxic growth in escherichia coli," *Biophys J*, vol. 83, no. 3, pp. 1331–40, 2002.
- [65] R. Caspi, H. Foerster, C. A. Fulcher, P. Kaipa, M. Krummenacker, M. Latendresse, S. Paley, S. Y. Rhee, A. G. Shearer, C. Tissier, *et al.*, "The metacyc database of metabolic pathways and enzymes and the biocyc collection of pathway/genome databases," *Nucleic acids research*, vol. 36, no. suppl_1, pp. D623–D631, 2007.

-
- [66] A. P. Arkin, R. W. Cottingham, C. S. Henry, N. L. Harris, R. L. Stevens, S. Maslov, P. Dehal, D. Ware, F. Perez, S. Canon, *et al.*, “Kbase: the united states department of energy systems biology knowledgebase,” *Nature Biotechnology*, vol. 36, no. 7, 2018.
- [67] A. Fredrickson, “Formulation of structured growth models,” *Biotechnology and bioengineering*, vol. 18, no. 10, pp. 1481–1486, 1976.
- [68] R. Milo, P. Jorgensen, U. Moran, G. Weber, and M. Springer, “Bionumbers—the database of key numbers in molecular and cell biology,” *Nucleic Acids Res*, vol. 38, no. Database issue, pp. D750–3, 2010.
- [69] H. Bremer and P. P. Dennis, “Modulation of chemical composition and other parameters of the cell by growth rate,” *Escherichia coli and Salmonella: cellular and molecular biology*, vol. 2, no. 2, pp. 1553–69, 1996.
- [70] B. S. Schuwirth, M. A. Borovinskaya, C. W. Hau, W. Zhang, A. Vila-Sanjurjo, J. M. Holton, and J. H. Cate, “Structures of the bacterial ribosome at 3.5 Å resolution,” *Science*, vol. 310, no. 5749, pp. 827–34, 2005.
- [71] J. Zhu, P. A. Penczek, R. Schroder, and J. Frank, “Three-dimensional reconstruction with contrast transfer function correction from energy-filtered cryoelectron micrographs: procedure and application to the 70S *Escherichia coli* ribosome,” *J Struct Biol*, vol. 118, no. 3, pp. 197–219, 1997.
- [72] R. Gilbert, “Physical biology of the cell, by rob phillips, jane kondev and julie theriot,” 2009.
- [73] F. C. Neidhardt, *The regulation of RNA synthesis in bacteria*, vol. 3, pp. 145–181. Elsevier, 1964.
- [74] S. H. J. Chan, M. N. Simons, and C. D. Maranas, “Steadycom: Predicting microbial abundances while ensuring community stability,” *PLoS computational biology*, vol. 13, no. 5, p. e1005539, 2017.
- [75] C. C. Petersen, “A note on transforming the product of variables to linear form in linear programs,” *Diskussionspapier, Purdue University*, 1971.
- [76] F. Glover, “Improved linear integer programming formulations of nonlinear integer problems,” *Management Science*, vol. 22, no. 4, pp. 455–460, 1975.
- [77] V. Hatzimanikatis, C. A. Floudas, and J. E. Bailey, “Analysis and design of metabolic reaction networks via mixed-integer linear optimization,” *AIChE Journal*, vol. 42, no. 5, pp. 1277–1292, 1996.
- [78] V. Hatzimanikatis, C. A. Floudas, and J. E. Bailey, “Optimization of regulatory architectures in metabolic reaction networks,” *Biotechnology and bioengineering*, vol. 52, no. 4, pp. 485–500, 1996.

Bibliography

- [79] J. Pramanik and J. Keasling, “Stoichiometric model of escherichia coli metabolism: incorporation of growth-rate dependent biomass composition and mechanistic energy requirements,” *Biotechnology and bioengineering*, vol. 56, no. 4, pp. 398–421, 1997.
- [80] Z. Kelman and M. O’Donnell, “Dna polymerase iii holoenzyme: structure and function of a chromosomal replicating machine,” *Annual review of biochemistry*, vol. 64, no. 1, pp. 171–200, 1995.
- [81] S. W. Peretti and J. E. Bailey, “Mechanistically detailed model of cellular metabolism for glucose-limited growth of escherichia coli b/r-a,” *Biotechnology and bioengineering*, vol. 28, no. 11, pp. 1672–1689, 1986.
- [82] S. W. Peretti and J. E. Bailey, “Simulations of host–plasmid interactions in escherichia coli: Copy number, promoter strength, and ribosome binding site strength effects on metabolic activity and plasmid gene expression,” *Biotechnology and bioengineering*, vol. 29, no. 3, pp. 316–328, 1987.
- [83] C. P. Selby, R. Drapkin, D. Reinberg, and A. Sancar, “Rna polymerase ii stalled at a thymine dimer: footprint and effect on excision repair,” *Nucleic acids research*, vol. 25, no. 4, pp. 787–793, 1997.
- [84] T. Benyamini, O. Folger, E. Ruppin, and T. Shlomi, “Flux balance analysis accounting for metabolite dilution,” *Genome biology*, vol. 11, no. 4, p. R43, 2010.
- [85] G.-W. Li, O. G. Berg, and J. Elf, “Effects of macromolecular crowding and dna looping on gene regulation kinetics,” *Nature Physics*, vol. 5, no. 4, pp. 294–297, 2009.
- [86] L. Bintu, N. E. Buchler, H. G. Garcia, U. Gerland, T. Hwa, J. Kondev, and R. Phillips, “Transcriptional regulation by the numbers: models,” *Current opinion in genetics & development*, vol. 15, no. 2, pp. 116–124, 2005.
- [87] M. D. Jankowski, C. S. Henry, L. J. Broadbelt, and V. Hatzimanikatis, “Group contribution method for thermodynamic analysis of complex metabolic networks,” *Biophysical Journal*, vol. 95, no. 3, pp. 1487–1499, 2008.
- [88] J. A. Bernstein, A. B. Khodursky, P. H. Lin, S. Lin-Chao, and S. N. Cohen, “Global analysis of mrna decay and abundance in escherichia coli at single-gene resolution using two-color fluorescent dna microarrays,” *Proceedings of the National Academy of Sciences of the United States of America*, vol. 99, no. 15, pp. 9697–9702, 2002.
- [89] M. A. Moran, B. Satinsky, S. M. Gifford, H. Luo, A. Rivers, L.-K. Chan, J. Meng, B. P. Durham, C. Shen, V. A. Varaljay, *et al.*, “Sizing up metatranscriptomics,” *The ISME journal*, vol. 7, no. 2, p. 237, 2013.
- [90] D. Davidi, E. Noor, W. Liebermeister, A. Bar-Even, A. Flamholz, K. Tummler, U. Barenholz, M. Goldenfeld, T. Shlomi, and R. Milo, “Global characterization of

- in vivo enzyme catalytic rates and their correspondence to in vitro $k(\text{cat})$ measurements,” *Proceedings of the National Academy of Sciences of the United States of America*, vol. 113, no. 12, pp. 3401–3406, 2016.
- [91] Z. A. King, A. Drager, A. Ebrahim, N. Sonnenschein, N. E. Lewis, and B. O. Palsson, “Escher: A web application for building, sharing, and embedding data-rich visualizations of biological pathways,” *Plos Computational Biology*, vol. 11, no. 8, 2015.
- [92] I. M. Keseler, J. Collado-Vides, S. Gama-Castro, J. Ingraham, S. Paley, I. T. Paulsen, M. Peralta-Gil, and P. D. Karp, “Ecocyc: a comprehensive database resource for *escherichia coli*,” *Nucleic acids research*, vol. 33, no. suppl_1, pp. D334–D337, 2005.
- [93] M. Travers, S. M. Paley, J. Shrager, T. A. Holland, and P. D. Karp, “Groups: knowledge spreadsheets for symbolic biocomputing,” *Database*, vol. 2013, 2013.
- [94] P. Salvy, G. Fengos, M. Ataman, T. Pathier, K. C. Soh, and V. Hatzimanikatis, “pytfa and mattfa: A python package and a matlab toolbox for thermodynamics-based flux analysis,” *Bioinformatics*, 2018.
- [95] A. Ebrahim, J. A. Lerman, B. O. Palsson, and D. R. Hyduke, “Cobrapy: Constraints-based reconstruction and analysis for python,” *Bmc Systems Biology*, vol. 7, 2013.
- [96] K. Jensen, J. Cardoso, and N. Sonnenschein, “Optlang: An algebraic modeling language for mathematical optimization,” *Journal of Open Source Software*, 2016.
- [97] A. Dalke, B. Wilczynski, B. A. Chapman, C. J. Cox, F. Kauff, I. Friedberg, J. T. Chang, M. J. L. de Hoon, P. J. A. Cock, T. Hamelryck, and T. Antao, “Biopython: freely available Python tools for computational molecular biology and bioinformatics,” *Bioinformatics*, vol. 25, pp. 1422–1423, 03 2009.
- [98] P. Salvy and V. Hatzimanikatis, “Emergence of diauxie as an optimal growth strategy under resource allocation constraints in cellular metabolism,” *bioRxiv*, 2020.
- [99] J. Monod, “The growth of bacterial cultures,” *Annual review of microbiology*, vol. 3, no. 1, pp. 371–394, 1949.
- [100] A. Ullmann, “Catabolite repression: a story without end,” *Research in microbiology*, vol. 147, no. 6-7, pp. 455–458, 1996.
- [101] A. Kremling, J. Geiselman, D. Ropers, and H. de Jong, “Understanding carbon catabolite repression in *escherichia coli* using quantitative models,” *Trends in microbiology*, vol. 23, no. 2, pp. 99–109, 2015.
- [102] F. Jacob, A. Ullmann, and J. Monod, “Délétions fusionnant l’opéron lactose et un opéron purine chez *escherichia coli*,” *Journal of Molecular Biology*, vol. 13, no. 3, pp. 704–719, 1965.

Bibliography

- [103] A. J. Griffiths, W. M. Gelbart, R. C. Lewontin, and J. H. Miller, *Modern genetic analysis: integrating genes and genomes*, vol. 1. Macmillan, 2002.
- [104] M. Durot, P.-Y. Bourguignon, and V. Schachter, “Genome-scale models of bacterial metabolism: reconstruction and applications,” *FEMS microbiology reviews*, vol. 33, no. 1, pp. 164–190, 2008.
- [105] E. J. O’Brien, J. M. Monk, and B. O. Palsson, “Using genome-scale models to predict biological capabilities,” *Cell*, vol. 161, no. 5, pp. 971–987, 2015.
- [106] B. Magasanik, “Catabolite repression,” in *Cold Spring Harbor symposia on quantitative biology*, vol. 26, pp. 249–256, Cold Spring Harbor Laboratory Press, 1961.
- [107] L. Yang, A. Ebrahim, C. J. Lloyd, M. A. Saunders, and B. O. Palsson, “Dynamicme: dynamic simulation and refinement of integrated models of metabolism and protein expression,” *BMC Systems Biology*, vol. 13, p. 2, Jan 2019.
- [108] M. Muir, L. Williams, and T. Ferenci, “Influence of transport energization on the growth yield of escherichia coli,” *Journal of bacteriology*, vol. 163, no. 3, pp. 1237–1242, 1985.
- [109] L. F. Leloir, “The enzymatic transformation of uridine diphosphate glucose into a galactose derivative,” *Archives of biochemistry and biophysics*, vol. 33, no. 2, pp. 186–190, 1951.
- [110] E. S. Maxwell, K. Kurahashi, and H. M. Kalckar, “Enzymes of the leloir pathway,” in *Methods in enzymology*, vol. 5, pp. 174–189, Elsevier, 1962.
- [111] C. A. Sellick, R. N. Campbell, and R. J. Reece, “Galactose metabolism in yeast—structure and regulation of the leloir pathway enzymes and the genes encoding them,” *International review of cell and molecular biology*, vol. 269, pp. 111–150, 2008.
- [112] B. Enjalbert, M. Coccagn-Bousquet, J.-C. Portais, and F. Letisse, “Acetate exposure determines the diauxic behavior of escherichia coli during the glucose-acetate transition,” *Journal of bacteriology*, vol. 197, no. 19, pp. 3173–3181, 2015.
- [113] A. Succurro, D. Segre, and O. Ebenhöf, “Emergent subpopulation behavior uncovered with a community dynamic metabolic model of escherichia coli diauxic growth,” *MSystems*, vol. 4, no. 1, pp. e00230–18, 2019.
- [114] W. F. Loomis and B. Magasanik, “Glucose-lactose diauxie in escherichia coli,” *Journal of bacteriology*, vol. 93, no. 4, pp. 1397–1401, 1967.
- [115] A. Kremling, K. Bettenbrock, B. Laube, K. Jahreis, J. Lengeler, and E. Gilles, “The organization of metabolic reaction networks: Iii. application for diauxic growth on glucose and lactose,” *Metabolic engineering*, vol. 3, no. 4, pp. 362–379, 2001.

-
- [116] N. Hadadi and V. Hatzimanikatis, "Design of computational retrobiosynthesis tools for the design of de novo synthetic pathways," *Current opinion in chemical biology*, vol. 28, pp. 99–104, 2015.
- [117] M. Ataman, D. F. H. Gardiol, G. Fengos, and V. Hatzimanikatis, "redgem: Systematic reduction and analysis of genome-scale metabolic reconstructions for development of consistent core metabolic models," *Plos Computational Biology*, vol. 13, no. 7, 2017.
- [118] M. Ataman and V. Hatzimanikatis, "lumpgem: Systematic generation of subnetworks and elementally balanced lumped reactions for the biosynthesis of target metabolites," *PLoS Comput Biol*, vol. 13, no. 7, p. e1005513, 2017.
- [119] B. Magasanik, *Glucose effects: inducer exclusion and repression*, pp. 189–219. Cold Spring Harbor Laboratory, 1970.
- [120] E. Perrin, V. Ghini, M. Giovannini, F. Di Patti, B. Cardazzo, L. Carraro, C. Fagorzi, P. Turano, R. Fani, and M. Fondi, "Diauxie and co-utilization of carbon sources can coexist during bacterial growth in nutritionally complex environments," *Nature communications*, vol. 11, no. 1, pp. 1–16, 2020.
- [121] J. C. Butcher and N. Goodwin, *Numerical methods for ordinary differential equations*, vol. 2. Wiley Online Library, 2008.
- [122] S. Boyd and L. Vandenberghe, *Convex optimization*. Cambridge university press, 2004.
- [123] S. Olsen and R. Brooker, "Analysis of the structural specificity of the lactose permease toward sugars," *Journal of Biological Chemistry*, vol. 264, no. 27, pp. 15982–15987, 1989.
- [124] D. D. Axe and J. E. Bailey, "Transport of lactate and acetate through the energized cytoplasmic membrane of escherichia coli," *Biotechnology and bioengineering*, vol. 47, pp. 8–19, 1995.
- [125] I. Borodina and J. Nielsen, "Advances in metabolic engineering of yeast *saccharomyces cerevisiae* for production of chemicals," *Biotechnology Journal*, vol. 9, no. 5, pp. 609–620, 2014.
- [126] A. Krivoruchko and J. Nielsen, "Production of natural products through metabolic engineering of *saccharomyces cerevisiae*," *Current opinion in biotechnology*, vol. 35, pp. 7–15, 2015.
- [127] T. Satyanarayana and G. Kunze, *Yeast diversity in human welfare*. Springer, 2017.
- [128] J. Forster, I. Famili, P. Fu, B. O. Palsson, and J. Nielsen, "Genome-scale reconstruction of the *saccharomyces cerevisiae* metabolic network," *Genome Research*, vol. 13, no. 2, pp. 244–253, 2003.

Bibliography

- [129] B. D. Heavner, K. Smallbone, B. Barker, P. Mendes, and L. P. Walker, “Yeast 5—an expanded reconstruction of the *saccharomyces cerevisiae* metabolic network,” *BMC systems biology*, vol. 6, no. 1, p. 55, 2012.
- [130] B. D. Heavner, K. Smallbone, N. D. Price, and L. P. Walker, “Version 6 of the consensus yeast metabolic network refines biochemical coverage and improves model performance,” *Database*, vol. 2013, 2013.
- [131] H. W. Aung, S. A. Henry, and L. P. Walker, “Revising the representation of fatty acid, glycerolipid, and glycerophospholipid metabolism in the consensus model of yeast metabolism,” *Ind Biotechnol (New Rochelle N Y)*, vol. 9, no. 4, pp. 215–228, 2013.
- [132] R. Chowdhury, A. Chowdhury, and C. D. Maranas, “Using gene essentiality and synthetic lethality information to correct yeast and cho cell genome-scale models,” *Metabolites*, vol. 5, no. 4, pp. 536–570, 2015.
- [133] H. Lu, F. Li, B. J. Sánchez, Z. Zhu, G. Li, I. Domenzain, S. Marcišauskas, P. M. Anton, D. Lappa, C. Lieven, *et al.*, “A consensus *s. cerevisiae* metabolic model yeast8 and its ecosystem for comprehensively probing cellular metabolism,” *Nature communications*, vol. 10, no. 1, pp. 1–13, 2019.
- [134] D. H. De Groot, J. Lischke, R. Muolo, R. Planqué, F. J. Bruggeman, and B. Teusink, “The common message of constraint-based optimization approaches: overflow metabolism is caused by two growth-limiting constraints,” *Cellular and Molecular Life Sciences*, pp. 1–13, 2019.
- [135] P. Van Hoek, J. P. Van Dijken, and J. T. Pronk, “Effect of specific growth rate on fermentative capacity of baker’s yeast,” *Applied and Environmental Microbiology*, vol. 64, no. 11, pp. 4226–4233, 1998.
- [136] L. G. Boender, E. A. de Hulster, A. J. van Maris, P. A. Daran-Lapujade, and J. T. Pronk, “Quantitative physiology of *saccharomyces cerevisiae* at near-zero specific growth rates,” *Applied and environmental microbiology*, vol. 75, no. 17, pp. 5607–5614, 2009.
- [137] K. Kasemets, I. Nisamedtinov, T.-M. Laht, K. Abner, and T. Paalme, “Growth characteristics of *saccharomyces cerevisiae* s288c in changing environmental conditions: auxo-accelerostat study,” *Antonie Van Leeuwenhoek*, vol. 92, no. 1, pp. 109–128, 2007.
- [138] T. V. Karpinets, D. J. Greenwood, C. E. Sams, and J. T. Ammons, “Rna: protein ratio of the unicellular organism as a characteristic of phosphorous and nitrogen stoichiometry and of the cellular requirement of ribosomes for protein synthesis,” *BMC biology*, vol. 4, no. 1, p. 30, 2006.

-
- [139] M. G. Vander Heiden, L. C. Cantley, and C. B. Thompson, "Understanding the warburg effect: the metabolic requirements of cell proliferation," *science*, vol. 324, no. 5930, pp. 1029–1033, 2009.
- [140] B. Xu, M. Jahic, and S.-O. Enfors, "Modeling of overflow metabolism in batch and fed-batch cultures of *escherichiacoli*," *Biotechnology progress*, vol. 15, no. 1, pp. 81–90, 1999.
- [141] J. Y. Kim, Y.-G. Kim, and G. M. Lee, "Cho cells in biotechnology for production of recombinant proteins: current state and further potential," *Applied microbiology and biotechnology*, vol. 93, no. 3, pp. 917–930, 2012.
- [142] A. Wagner, C. Wang, D. DeTomaso, J. Avila-Pacheco, S. Zaghouani, J. Fessler, S. Eyzaguirre, E. Akama-Garren, K. Pierce, N. Ron-Harel, *et al.*, "In silico modeling of metabolic state in single th17 cells reveals novel regulators of inflammation and autoimmunity," *bioRxiv*, 2020.
- [143] S. Hahn, "Structure and mechanism of the rna polymerase ii transcription machinery," *Nature structural & molecular biology*, vol. 11, no. 5, pp. 394–403, 2004.
- [144] B. Alberts, "Molecular biology of the cell," 2018.
- [145] R. J. Planta and W. H. Mager, "The list of cytoplasmic ribosomal proteins of *saccharomyces cerevisiae*," *Yeast (Chichester, England)*, vol. 14, no. 5, p. 471, 1998.
- [146] H.-R. Graack and B. Wittmann-Liebold, "Mitochondrial ribosomal proteins (mrps) of yeast," *Biochemical Journal*, vol. 329, no. 3, pp. 433–448, 1998.
- [147] M. L. Mavrouniotis, "Group contributions for estimating standard gibbs energies of formation of biochemical compounds in aqueous solution," *Biotechnology and bioengineering*, vol. 36, no. 10, pp. 1070–1082, 1990.
- [148] R. A. Alberty, *Thermodynamics of biochemical reactions*. John Wiley & Sons, 2005.
- [149] M. Kanehisa and S. Goto, "Kegg: kyoto encyclopedia of genes and genomes," *Nucleic acids research*, vol. 28, no. 1, pp. 27–30, 2000.
- [150] B. H. Meldal, O. Forner-Martinez, M. C. Costanzo, J. Dana, J. Demeter, M. Dumousseau, S. S. Dwight, A. Gaulton, L. Licata, A. N. Melidoni, *et al.*, "The complex portal-an encyclopaedia of macromolecular complexes," *Nucleic acids research*, vol. 43, no. D1, pp. D479–D484, 2015.
- [151] Q. Yuan, T. Huang, P. Li, T. Hao, F. Li, H. Ma, Z. Wang, X. Zhao, T. Chen, and I. Goryanin, "Pathway-consensus approach to metabolic network reconstruction for *pseudomonas putida* kt2440 by systematic comparison of published models," *PloS one*, vol. 12, no. 1, p. e0169437, 2017.

Bibliography

- [152] S. H. Chan, J. Cai, L. Wang, M. N. Simons-Senftle, and C. D. Maranas, “Standardizing biomass reactions and ensuring complete mass balance in genome-scale metabolic models,” *Bioinformatics*, vol. 33, no. 22, pp. 3603–3609, 2017.
- [153] H. Lange and J. Heijnen, “Statistical reconciliation of the elemental and molecular biomass composition of *saccharomyces cerevisiae*,” *Biotechnology and bioengineering*, vol. 75, no. 3, pp. 334–344, 2001.
- [154] A. K. Gombert, M. M. dos Santos, B. Christensen, and J. Nielsen, “Network identification and flux quantification in the central metabolism of *saccharomyces cerevisiae* under different conditions of glucose repression,” *Journal of bacteriology*, vol. 183, no. 4, pp. 1441–1451, 2001.
- [155] M. Wang, M. Weiss, M. Simonovic, G. Haertinger, S. P. Schrimpf, M. O. Hengartner, and C. von Mering, “Paxdb, a database of protein abundance averages across all three domains of life,” *Molecular & cellular proteomics*, vol. 11, no. 8, pp. 492–500, 2012.
- [156] I. Thiele and B. Ø. Palsson, “A protocol for generating a high-quality genome-scale metabolic reconstruction,” *Nature protocols*, vol. 5, no. 1, p. 93, 2010.
- [157] N. E. Lewis, K. K. Hixson, T. M. Conrad, J. A. Lerman, P. Charusanti, A. D. Polpitiya, J. N. Adkins, G. Schramm, S. O. Purvine, D. Lopez-Ferrer, *et al.*, “Omic data from evolved *e. coli* are consistent with computed optimal growth from genome-scale models,” *Molecular systems biology*, vol. 6, no. 1, p. 390, 2010.
- [158] M. Masid, M. Ataman, and V. Hatzimanikatis, “Analysis of human metabolism by reducing the complexity of the genome-scale models using redhuman,” *Nature Communications*, vol. 11, no. 1, pp. 1–12, 2020.
- [159] M. Masid, *Modeling metabolic and signaling pathways in cancer cells*. PhD thesis, École Polytechnique Fédérale de Lausanne, Lausanne, 2020.
- [160] M. Masid and V. Hatzimanikatis, “Model-based data integration and minimal network enrichment analysis identifies metabolic differences across cancer types.” 2020.
- [161] C. Holohan, S. Van Schaeybroeck, D. B. Longley, and P. G. Johnston, “Cancer drug resistance: an evolving paradigm,” *Nature Reviews Cancer*, vol. 13, no. 10, pp. 714–726, 2013.
- [162] S. Bhattacharyya, S. Saha, K. Giri, I. R. Lanza, K. S. Nair, N. B. Jennings, C. Rodriguez-Aguayo, G. Lopez-Berestein, E. Basal, A. L. Weaver, *et al.*, “Cystathionine beta-synthase (*cbs*) contributes to advanced ovarian cancer progression and drug resistance,” *PloS one*, vol. 8, no. 11, p. e79167, 2013.

-
- [163] S. D. Finley, L.-H. Chu, and A. S. Popel, “Computational systems biology approaches to anti-angiogenic cancer therapeutics,” *Drug discovery today*, vol. 20, no. 2, pp. 187–197, 2015.
- [164] D. Hanahan and R. A. Weinberg, “Hallmarks of cancer: the next generation,” *cell*, vol. 144, no. 5, pp. 646–674, 2011.
- [165] M. Franciosi, G. Lucisano, E. Lapice, G. F. Strippoli, F. Pellegrini, and A. Nicolucci, “Metformin therapy and risk of cancer in patients with type 2 diabetes: systematic review,” *PloS one*, vol. 8, no. 8, p. e71583, 2013.
- [166] R. Mamtani, N. Pfanzelter, K. Haynes, B. S. Finkelman, X. Wang, S. M. Keefe, N. B. Haas, D. J. Vaughn, and J. D. Lewis, “Incidence of bladder cancer in patients with type 2 diabetes treated with metformin or sulfonylureas,” *Diabetes care*, vol. 37, no. 7, pp. 1910–1917, 2014.
- [167] G. Rena, D. G. Hardie, and E. R. Pearson, “The mechanisms of action of metformin,” *Diabetologia*, vol. 60, no. 9, pp. 1577–1585, 2017.
- [168] M. Jain, R. Nilsson, S. Sharma, N. Madhusudhan, T. Kitami, A. L. Souza, R. Kafri, M. W. Kirschner, C. B. Clish, and V. K. Mootha, “Metabolite profiling identifies a key role for glycine in rapid cancer cell proliferation,” *Science*, vol. 336, no. 6084, pp. 1040–1044, 2012.
- [169] E. Brunk, S. Sahoo, D. C. Zielinski, A. Altunkaya, A. Dräger, N. Mih, F. Gatto, A. Nilsson, G. A. P. Gonzalez, M. K. Aurich, *et al.*, “Recon3d enables a three-dimensional view of gene variation in human metabolism,” *Nature biotechnology*, vol. 36, no. 3, p. 272, 2018.
- [170] T. Zhang, P. Guo, Y. Zhang, H. Xiong, X. Yu, S. Xu, X. Wang, D. He, and X. Jin, “The antidiabetic drug metformin inhibits the proliferation of bladder cancer cells in vitro and in vivo,” *International journal of molecular sciences*, vol. 14, no. 12, pp. 24603–24618, 2013.
- [171] R. DeFronzo, G. A. Fleming, K. Chen, and T. A. Bicsak, “Metformin-associated lactic acidosis: Current perspectives on causes and risk,” *Metabolism*, vol. 65, no. 2, pp. 20–29, 2016.
- [172] F. Klemm, R. R. Maas, R. L. Bowman, M. Kornete, K. Soukup, S. Nassiri, J.-P. Brouland, C. A. Iacobuzio-Donahue, C. Brennan, V. Tabar, *et al.*, “Interrogation of the microenvironmental landscape in brain tumors reveals disease-specific alterations of immune cells,” *Cell*, 2020.
- [173] D. Li and S. D. Finley, “The impact of tumor receptor heterogeneity on the response to anti-angiogenic cancer treatment,” *Integrative Biology*, vol. 10, no. 4, pp. 253–269, 2018.

Bibliography

- [174] G. L. Szeto and S. D. Finley, "Integrative approaches to cancer immunotherapy," *Trends in cancer*, vol. 5, no. 7, pp. 400–410, 2019.
- [175] E. M. Malinovskaya, E. S. Ershova, V. E. Golimbet, L. N. Porokhovnik, N. A. Lyapunova, S. I. Kutsev, N. N. Veiko, and S. V. Kostyuk, "Copy number of human ribosomal genes with aging: unchanged mean, but narrowed range and decreased variance in elderly group," *Frontiers in genetics*, vol. 9, p. 306, 2018.
- [176] P. Romero, J. Wagg, M. L. Green, D. Kaiser, M. Krummenacker, and P. D. Karp, "Computational prediction of human metabolic pathways from the complete human genome," *Genome biology*, vol. 6, no. 1, p. R2, 2005.
- [177] J. Y. Ryu, H. U. Kim, and S. Y. Lee, "Framework and resource for more than 11,000 gene-transcript-protein-reaction associations in human metabolism," *Proceedings of the National Academy of Sciences*, vol. 114, no. 45, pp. E9740–E9749, 2017.
- [178] T. M. Thomson, K. R. Benjamin, A. Bush, T. Love, D. Pincus, O. Resnekov, C. Y. Richard, A. Gordon, A. Colman-Lerner, D. Endy, *et al.*, "Scaffold number in yeast signaling system sets tradeoff between system output and dynamic range," *Proceedings of the National Academy of Sciences*, vol. 108, no. 50, pp. 20265–20270, 2011.
- [179] H. Pertoft and T. C. Laurent, "Isopycnic separation of cells and cell organelles by centrifugation in modified colloidal silica gradients," in *Methods of cell separation*, pp. 25–65, Springer, 1977.
- [180] J. G. Fedor, A. J. Jones, A. Di Luca, V. R. Kaila, and J. Hirst, "Correlating kinetic and structural data on ubiquinone binding and reduction by respiratory complex i," *Proceedings of the National Academy of Sciences*, vol. 114, no. 48, pp. 12737–12742, 2017.
- [181] E. Fontaine, "Metformin-induced mitochondrial complex i inhibition: facts, uncertainties, and consequences," *Frontiers in endocrinology*, vol. 9, p. 753, 2018.
- [182] T. W. Traut, *Allosteric regulatory enzymes*, p. 222. Springer Science & Business Media, 2007.
- [183] J. M. Corton, J. G. Gillespie, S. A. Hawley, and D. G. Hardie, "5-aminoimidazole-4-carboxamide ribonucleoside: a specific method for activating amp-activated protein kinase in intact cells?," *European journal of biochemistry*, vol. 229, no. 2, pp. 558–565, 1995.
- [184] J. Weekes, S. A. Hawley, J. Corton, D. Shugar, and D. G. Hardie, "Activation of rat liver amp-activated protein kinase by kinase kinase in a purified, reconstituted system: Effects of amp and amp analogues," *European journal of biochemistry*, vol. 219, no. 3, pp. 751–757, 1994.

-
- [185] M. Escós, P. Latorre, J. Hidalgo, R. Hurtado-Guerrero, J. A. Carrodegua, and P. López-Buesa, “Kinetic and functional properties of human mitochondrial phosphoenolpyruvate carboxykinase,” *Biochemistry and biophysics reports*, vol. 7, pp. 124–129, 2016.
- [186] M. Ataman and V. Hatzimanikatis, “Heading in the right direction: thermodynamics-based network analysis and pathway engineering,” *Current Opinion in Biotechnology*, vol. 36, pp. 176–182, 2015.
- [187] K. C. Soh and V. Hatzimanikatis, “Network thermodynamics in the post-genomic era,” *Current Opinion in Microbiology*, vol. 13, no. 3, pp. 350–357, 2010.
- [188] R. M. T. Fleming and I. Thiele, “von bertalanffy 1.0: a cobra toolbox extension to thermodynamically constrain metabolic models,” *Bioinformatics*, vol. 27, no. 1, pp. 142–143, 2011.
- [189] R. M. T. Fleming, I. Thiele, and H. P. Nasheuer, “Quantitative assignment of reaction directionality in constraint-based models of metabolism: Application to escherichia coli,” *Biophysical Chemistry*, vol. 145, no. 2-3, pp. 47–56, 2009.
- [190] N. Zamboni, A. Kummel, and M. Heinemann, “anet: a tool for network-embedded thermodynamic analysis of quantitative metabolome data,” *Bmc Bioinformatics*, vol. 9, 2008.
- [191] S. Andreozzi, A. Chakrabarti, K. C. Soh, A. Burgard, T. H. Yang, S. Van Dien, L. Miskovic, and V. Hatzimanikatis, “Identification of metabolic engineering targets for the enhancement of 1,4-butanediol production in recombinant e. coli using large-scale kinetic models,” *Metab Eng*, vol. 35, pp. 148–159, 2016.
- [192] A. Chakrabarti, L. Miskovic, K. C. Soh, and V. Hatzimanikatis, “Towards kinetic modeling of genome-scale metabolic networks without sacrificing stoichiometric, thermodynamic and physiological constraints,” *Biotechnology Journal*, vol. 8, no. 9, pp. 1043–U105, 2013.
- [193] A. Chiappino-Pepe, S. Tymoshenko, M. Ataman, D. Soldati-Favre, and V. Hatzimanikatis, “Bioenergetics-based modeling of plasmodium falciparum metabolism reveals its essential genes, nutritional requirements, and thermodynamic bottlenecks,” *Plos Computational Biology*, vol. 13, no. 3, 2017.
- [194] A. Kiparissides and V. Hatzimanikatis, “Thermodynamics-based metabolite sensitivity analysis in metabolic networks,” *Metabolic Engineering*, vol. 39, pp. 117–127, 2017.
- [195] L. Miskovic, S. Alff-Tuomala, K. C. Soh, D. Barth, L. Salusjärvi, J.-P. Pitkänen, L. Ruohonen, M. Penttilä, and V. Hatzimanikatis, “A design–build–test cycle using modeling and experiments reveals interdependencies between upper glycolysis and

Bibliography

- xylose uptake in recombinant *s. cerevisiae* and improves predictive capabilities of large-scale kinetic models,” *Biotechnology for Biofuels*, vol. 10, no. 1, p. 166, 2017.
- [196] G. Savoglidis, A. X. D. dos Santos, I. Riezman, P. Angelino, H. Riezman, and V. Hatzimanikatis, “A method for analysis and design of metabolism using metabolomics data and kinetic models: Application on lipidomics using a novel kinetic model of sphingolipid metabolism,” *Metabolic Engineering*, vol. 37, pp. 46–62, 2016.
- [197] P. Debye and E. Hückel, “Zur theorie der elektrolyte. i. gefrierpunktserniedrigung und verwandte erscheinungen. the theory of electrolytes. i. lowering of freezing point and related phenomena,” *Physikalische Zeitschrift*, vol. 24, pp. 185–206, 1923.
- [198] J. Szegezdi and F. Csizmadia, “Method for calculating the pka values of small and large molecules,” in *ABSTRACTS OF PAPERS OF THE AMERICAN CHEMICAL SOCIETY*, vol. 233, AMER CHEMICAL SOC 1155 16TH ST, NW, WASHINGTON, DC 20036 USA, 2007.
- [199] R. Overbeek, T. Begley, R. M. Butler, J. V. Choudhuri, H.-Y. Chuang, M. Cohoon, V. de Crécy-Lagard, N. Diaz, T. Disz, and R. Edwards, “The subsystems approach to genome annotation and its use in the project to annotate 1000 genomes,” *Nucleic acids research*, vol. 33, no. 17, pp. 5691–5702, 2005.
- [200] P. Salvy, M. Ataman, G. Fengos, B. Mouscadet, R. Poirot, and V. Hatzimanikatis, “A python implementation of the metabolic network analysis and reduction algorithms redgem and lumpgem,” *bioRxiv*, 2020.
- [201] L. Miskovic, M. Tokic, G. Fengos, and V. Hatzimanikatis, “Rites of passage: requirements and standards for building kinetic models of metabolic phenotypes,” *Current Opinion in Biotechnology*, vol. 36, pp. 146–153, 2015.
- [202] L. Miskovic and V. Hatzimanikatis, “Production of biofuels and biochemicals: in need of an oracle,” *Trends in Biotechnology*, vol. 28, no. 8, pp. 391–397, 2010.
- [203] C. Chassagnole, N. Noisommit-Rizzi, J. W. Schmid, K. Mauch, and M. Reuss, “Dynamic modeling of the central carbon metabolism of *escherichia coli*,” *Biotechnology and bioengineering*, vol. 79, no. 1, pp. 53–73, 2002.
- [204] P. Erdrich, R. Steuer, and S. Klamt, “An algorithm for the reduction of genome-scale metabolic network models to meaningful core models,” *BMC systems biology*, vol. 9, no. 1, p. 48, 2015.
- [205] B. Teusink, J. Passarge, C. A. Reijenga, E. Esgalhado, C. C. Van der Weijden, M. Schepper, M. C. Walsh, B. M. Bakker, K. Van Dam, H. V. Westerhoff, *et al.*, “Can yeast glycolysis be understood in terms of in vitro kinetics of the constituent enzymes? testing biochemistry,” *European Journal of Biochemistry*, vol. 267, no. 17, pp. 5313–5329, 2000.

-
- [206] T. Hameri, G. Fengos, M. Ataman, L. Miskovic, and V. Hatzimanikatis, “Kinetic models of metabolism that consider alternative steady-state solutions of intracellular fluxes and concentrations,” *Metabolic engineering*, vol. 52, pp. 29–41, 2019.
- [207] M. Tokic, N. Hadadi, M. Ataman, D. Neves, B. E. Ebert, L. M. Blank, L. Miskovic, and V. Hatzimanikatis, “Discovery and evaluation of biosynthetic pathways for the production of five methyl ethyl ketone precursors,” *ACS synthetic biology*, vol. 7, no. 8, pp. 1858–1873, 2018.
- [208] D. R. Weilandt and V. Hatzimanikatis, “Particle-based simulation reveals macro-molecular crowding effects on the michaelis-menten mechanism,” *Biophysical journal*, vol. 117, no. 2, pp. 355–368, 2019.
- [209] L. C. Bryan, D. R. Weilandt, A. L. Bachmann, S. Kilic, C. C. Lechner, P. D. Odermatt, G. E. Fantner, S. Georgeon, O. Hantschel, V. Hatzimanikatis, *et al.*, “Single-molecule kinetic analysis of hp1-chromatin binding reveals a dynamic network of histone modification and dna interactions,” *Nucleic acids research*, vol. 45, no. 18, pp. 10504–10517, 2017.
- [210] D. R. Weilandt, *Modeling biochemical reaction networks in complex intracellular environments*. PhD thesis, École Polytechnique Fédérale de Lausanne, Lausanne, 2020.
- [211] A. Khodayari and C. D. Maranas, “A genome-scale escherichia coli kinetic metabolic model k-ecoli457 satisfying flux data for multiple mutant strains,” *Nature communications*, vol. 7, no. 1, pp. 1–12, 2016.
- [212] P. A. Saa and L. K. Nielsen, “Construction of feasible and accurate kinetic models of metabolism: a bayesian approach,” *Scientific reports*, vol. 6, no. 1, pp. 1–13, 2016.
- [213] E. V. Nikolaev, P. Pharkya, C. D. Maranas, and A. Armaou, “Optimal selection of enzyme levels using large-scale kinetic models,” *IFAC Proceedings Volumes*, vol. 38, no. 1, pp. 25–30, 2005.
- [214] M. Tokic, V. Hatzimanikatis, and L. Miskovic, “Large-scale kinetic metabolic models of pseudomonas putida kt2440 for consistent design of metabolic engineering strategies,” *Biotechnology for biofuels*, vol. 13, no. 1, pp. 1–19, 2020.
- [215] L. Q. Wang, I. Birol, and V. Hatzimanikatis, “Metabolic control analysis under uncertainty: Framework development and case studies,” *Biophysical Journal*, vol. 87, no. 6, pp. 3750–3763, 2004.
- [216] A. C. Hindmarsh, P. N. Brown, K. E. Grant, S. L. Lee, R. Serban, D. E. Shumaker, and C. S. Woodward, “Sundials: Suite of nonlinear and differential/algebraic equation solvers,” *ACM Transactions on Mathematical Software (TOMS)*, vol. 31, no. 3, pp. 363–396, 2005.

Bibliography

- [217] B. Malengier, P. Kison, J. Tocknell, C. Abert, F. Bruckner, and M.-A. Bisotti, “Odes: a high level interface to ode and dae solvers,” *J. Open Source Software*, vol. 3, no. 22, p. 165, 2018.
- [218] J. R. Karr, J. C. Sanghvi, D. N. Macklin, M. V. Gutschow, J. M. Jacobs, B. Bolival Jr, N. Assad-Garcia, J. I. Glass, and M. W. Covert, “A whole-cell computational model predicts phenotype from genotype,” *Cell*, vol. 150, no. 2, pp. 389–401, 2012.
- [219] D. N. Macklin, N. A. Ruggero, and M. W. Covert, “The future of whole-cell modeling,” *Current opinion in biotechnology*, vol. 28, pp. 111–115, 2014.
- [220] Z. A. King, C. J. Lloyd, A. M. Feist, and B. O. Palsson, “Next-generation genome-scale models for metabolic engineering,” *Current opinion in biotechnology*, vol. 35, pp. 23–29, 2015.
- [221] J. Mounier, C. Monnet, T. Vallaey, R. Arditi, A.-S. Sarthou, A. Hélias, and F. Irlinger, “Microbial interactions within a cheese microbial community,” *Applied and environmental microbiology*, vol. 74, no. 1, pp. 172–181, 2008.
- [222] Y. Liu, S. Rousseaux, R. Tourdot-Maréchal, M. Sadoudi, R. Gougeon, P. Schmitt-Kopplin, and H. Alexandre, “Wine microbiome: a dynamic world of microbial interactions,” *Critical reviews in food science and nutrition*, vol. 57, no. 4, pp. 856–873, 2017.
- [223] S. Blasche, Y. Kim, R. Mars, E. Kafkia, M. Maansson, D. Machado, B. Teusink, J. Nielsen, V. Benes, R. Neves, *et al.*, “Emergence of stable coexistence in a complex microbial community through metabolic cooperation and spatio-temporal niche partitioning,” *bioRxiv*, p. 541870, 2019.
- [224] E. Helms, M. K. Onate, and M. H. Sherman, “Fibroblast heterogeneity in the pancreatic tumor microenvironment,” *Cancer Discovery*, vol. 10, no. 5, pp. 648–656, 2020.
- [225] V. Hatzimanikatis, *Analysis and design of metabolic reaction networks*. PhD thesis, California Institute of Technology, 1997.

Glossary

Big-M value A value that is systematically bigger than the other variables in presence within an expression. Used with binary variables to model **if**-type logical dependencies in an optimization problem. Often annotated M in expressions.

Bilinear(ity) A function is said to be bilinear if it contains a product of two of its variables. This term is called a bilinearity. A problem with a constraint defined by a bilinear function of variables is said to be bilinear. That is the case in the non-linearized expression problem with the term $\mu * E_j$, where both μ and E_j are variables of the problem.

Binary variable An integer variable whose value is constrained to 0 or 1. Used to model **if**-type logical dependencies in an optimization problem. For instance, they are used in TFA to enforce the statement “if the Gibbs free energies of this reaction is negative, its net biochemical flux will be in its forward direction”. Inclusion of binary variables in a LP problem make it MILP.

Discretization Process by which a continuous variable is replaced by a set of representative discrete values it can take. We use it in ETFL to approximate μ and perform a linearization. Sampling is a type of discretization.

Eukaryote / eukaryotic organism Organisms whose genetic information is enclosed by a membrane in a nucleus, as opposed to prokaryotes.

Fluxomics (exo) Experimental data accounting for biochemical reaction fluxes. Ex-fluxomics are for fluxes outside the cell.

Genotype The genetic information carried by an organism’s DNA.

Glossary

Knock-out (Gene) The action of silencing the expression of a gene. The organism is then grown (or simulated) without the ability to express and use this specific gene. This can be done experimentally through targeted gene editing, and *in silico* by the suppression of adequate reactions in which the gene participates.

Linearization Process by which a non-linear function is approximated by a linear approximant. In the case of ETFL, we discretize μ to make the bilinear terms $\mu * E_j$ (piecewise-)linear.

LP Linear program. An optimization formulation where a problem is defined by a linear objective function, a set of linear equalities and a set of linear inequalities. FBA is a kind of LP.

Metabolomics (exo) Experimental data accounting for metabolites concentrations. Can be relative or absolute. Exometabolomics are for compounds outside the cell.

MILP A LP with integer variables. The problem is then piecewise-linear, and requires specific solving methods. When all the integer variables are fixed, a LP is obtained. TFA is a kind of MILP.

Omics General term to regroup several types of experimental data that can be gathered from cellular cultures. These includes metabolomics, proteomics, transcriptomics, and fluxomics.

Phenotype The observable traits of the cell, understood as a product of the information flow from the DNA to the enzymes controlling metabolism.

Prokaryote / prokaryotic organism Organisms whos egenetic information is not membrane-bound within the cytoplasm, as opposed to eukaryotes.

Proteomics Experimental data accounting for the concentrations of proteins in a cell. Can be relative or absolute amounts.

Special Ordered Set of type 1 (SOS1) constraint A type of constraint where a sum of binary variables has to be lower than or equal to 1. Useful to model a choice between different possibilities.

Transcriptomics Experimental data accounting for the concentrations of mRNAs in a cell. Can be relative or absolute amounts

Zeroth order approximation Approximation of a function using a piece-wise constant function. The values of the zeroth-order approximation of the function are a discretization of the space of values of the initial function.

Pierre Salvy

pierre.salvy@outlook.com | +41 78 669 53 79 | [Github] psalvy [LinkedIn] psalvy [Twitter] @psalvy_

EDUCATION

EPFL

PHD - COMPUTATIONAL BIOLOGY
Grad. Aug 2020 | Lausanne, Switzerland

Prof. Vassily Hatzimanikatis
Laboratory of Computational Systems Biotechnology
Marie Skłodowska-Curie Grant

MINES PARISTECH

MASTER OF SCIENCE AND EXECUTIVE ENGINEERING
Grad. Sep 2016 | Paris, France
Sp. Biotechnology

LYCEE STE GENEVIEVE

Grad. July 2010 | Versailles, France

COURSEWORK

GRADUATE

Convex Optimization
Operations Research • Data Analysis
Industrial Biotechnology
Quantum Physics • Statistical Physics
Nuclear Engineering • Material Science

UNDERGRADUATE

Calculus • Algebra
Control Theory (*Examiner 1+3x*)
Fluid Mechanics • Electromagnetism

SKILLS

PROGRAMMING

Languages
Python • Java • Matlab • Shell
MySQL • Dockerfile • \LaTeX
Packages
Pandas • NumPy • Scipy
Keras • PyTorch • TensorFlow

LANGUAGES

French (*Native*) • English (*Bilingual*)
Spanish (*Professional Capacity*)
Russian (*Beginner*)

ACHIEVEMENTS

2019 EPFL Excellence teacher
2012 9th/3575 Natl. Exam Mines-Ponts
2012 20th/4578 Natl. Exam E3A

INTERESTS

Rock Climbing • Skiing
Waterpolo (*Competitive*)
Analog Photography (*Portrait*)
Programming projects (*chatbot, data viz*)

PHD EXPERIENCE

METABOLISM AND EXPRESSION MODELS | MAIN RESEARCH

Nov 2016 – Sep 2020 | Lausanne, Switzerland
Developed a new framework for the formulation of ME-Models which also includes thermodynamics and resource allocation. Allows straightforward 'omics integration.

- Development of the first *E. coli* and *S. cerevisiae* thermodynamically-enabled ME-models
- Dynamic thermodynamically-enabled ME-models successfully predicting diauxic behavior in *E. coli*
- Cancer-specific Human models with transcriptomics integration

DEEP NNS FOR KINETIC MODELS | RESEARCH / STUDENT ADVISOR

Dec 2017 – Sep 2020 | Lausanne, Switzerland
Investigated the use of Generative Adversarial Networks for generating stable and condition-dependent parameter sets for large-scale kinetic models. The project benefited from the help of 5 master students I co-advised.

LABORATORY IT MANAGER | SUPPORT

Nov 2016 – Sep 2020 | Lausanne, Switzerland

- Managing the transition to open-source and open science
- Admin of 50 Git repos spanning 30 collaborators
- Organized and gave Git/Docker/general programming trainings

INDUSTRY EXPERIENCE

IDENTITY PURSUIT | DATA ADVISOR

Apr 2015 – Feb 2017 | Lausanne, Switzerland
Machine-learning algorithms to model user behavior in a psychometric application. Market research on competitors and ML technologies for behavioral inference.

TOTAL / AMYRIS | COMPUTATIONAL BIOLOGY INTERN

Apr – Oct 2015 | Emeryville, CA
Developed a ¹³C Metabolic Flux Analysis pipeline, integrated to production.

TOTAL / AMYRIS + EPFL | SCIENTIFIC COMPUTING INTERN

Aug 2014 – Oct 2015 | Emeryville, CA + Lausanne, Switzerland
S. cerevisiae genome-scale models to produce biofuels. Created a database-powered model generator to predict recombinant strain yields. Organized a tripartite collaboration with EPFL to model thermodynamics for better physiology predictions.

OTHER EXPERIENCE

DATA-DRIVEN DIAGNOSIS OF RARE METABOLIC DISEASES

Oct 2012 – Feb 2013 | Paris, France
Manager of a project aiming at improving the diagnosis of rare neuro-metabolic diseases, in cooperation with the Hospital de la Pitié-Salpêtrière (Paris V). Algorithms based on linear algebra and graph theory to help MDs diagnose patients.

PUBLICATIONS

- [1] P. Salvy and V. Hatzimanikatis, "The etfl formulation allows multi-omics integration in thermodynamics-compliant metabolism and expression models," *Nature Communications*, vol. 11, no. 1, pp. 1–17, 2020.
- [2] P. Salvy, G. Fengos, M. Ataman, T. Pathier, K. C. Soh, and V. Hatzimanikatis, "pytfa and mattfa: a python package and a matlab toolbox for thermodynamics-based flux analysis," *Bioinformatics*, vol. 35, no. 1, pp. 167–169, 2018.
- [3] P. Salvy and V. Hatzimanikatis, "Diauxie in *e. coli* is an optimal growth behavior under proteome limitation," *In Preparation*, 2020.
- [4] P. Salvy, M. Masid, and V. Hatzimanikatis, "Models of metabolism and expression reproduce the growth-inhibition effect of metformin on colon cancer cells," *In Preparation*, 2020.