EPFL

Thèse n° 7725

Modeling metabolic and signaling pathways in cancer cells

Présentée le 25 septembre 2020

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

pour l'obtention du grade de Docteur ès Sciences

par

Maria MASID BARCON

Acceptée sur proposition du jury

Prof. K. Sivula, président du jury
Prof. V. Hatzimanikatis, directeur de thèse
Dr B. Schoeberl, rapporteuse
Prof. I. Thiele, rapporteuse
Prof. D. Hanahan, rapporteur

École polytechnique fédérale de Lausanne

2020

To Mami and Lili …

… the two women that made me be who I am

Acknowledgements

Looking back through the last four years to write these acknowledgements, I have realized how many people have directly or indirectly contributed to the success of my PhD and of this thesis.

First, and foremost, I would like to express my deepest gratitude to my supervisor Prof. Vassily Hatzimanikatis. Vassily, you believed on me and gave me the unique opportunity of doing a PhD in your group. I am very grateful for all your advice, knowledge, support, and scientific and life learnings. Having you as an academic father and working with you all these years allowed me to grow as a scientist and as a person.

Secondly, I would like to thank the president and the committee members of my PhD defence. Prof. Kevin Sivula for being an amazing president and organizing all the thesis defence despite the situation. Prof. Douglas Hanahan, Dr. Birgit Schoeberl and Prof. Ines Thiele for reading my thesis, and for the inspiring scientific discussion, making the thesis defence a memorable experience. It is a great honour to have had the opportunity of discussing with you and had your insightful comments and constructive feedback.

I would also like to thank Christine Kupper, Anne Lene Odegaard, and Melody Meyer for their support and help with all the administrative work.

Next, I would like to thank my academic family: the LCSB! Without them, my personal and professional time in EPFL would not have been the same. A big big thank you goes to Misko, for being always willing to listen and help when it was needed. Thank you also to my co-workers and friends: Meriç, you were the best *mentor* I could have had, thank you for sharing all your knowledge with me and helping me every time I needed it; Georgios, thank you for all the coffee&lunch times, and for always seeing the positive side of things. To both of them, thank you for all the fun moments these years, it is a real pleasure to know you both; Vikash, thank you for introducing me to the field of gene expression and signaling, and for being so patient and helpful.

I would also like to express my gratitude to my best friends in the lab and in Lausanne: Daniel, Jasmin, Pierre, and Sophia, this thesis would not be possible without them. We have gone together through all the stages of the PhD. Sophia, my team, the first person I met when I arrived to Lausanne. Thank you so much for all the breaks, all the wines and all the talking! Daniel, thank you for being you! For all the support, coffees, beers, and discussions. Pierre, thank you for making out of the office such a fantastic place to be in every morning and for supporting me. And Jasmin, thank you for your calmness, honesty, and friendship. I would also like to thank Zhaleh and Homa, for being always so friendly and helpful. I could not forget the youngest, Evangelia and Asli, for all the socializing events. Tuure and Tiziano, for all the epic, unforgettable nights. Noushin, for your friendship and advice during my PhD. Finally, all the present and past members of LCSB: Georgios S., Liliana, Yves, Stepan, Milenko, Beatriz, Anastasia, Anush, Omid, Yiannis, Joana, Robin, Aarti, and Daniel H. for creating such a nice environment in the lab. And my students: Maxime, Diane, Mélanie, Elliott, and Samy, for giving me the opportunity of teaching and guiding them through their master thesis.

This thesis would not have been possible without the financial support of the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska Curie Grant Agreement No. 675585 SyMBioSys. Being a Marie Curie fellow has been a enjoyable and enriching experience. I want to thank the members of the SyMBioSys ITN, all the students and all the professors. Specially Mixalis and José for adopting me in their group and teaching me all about cells and experimental procedures during my stay in London. Nikos and Jake for making my stay in Vigo a great experience. And a special thanks to Cristina and Prof. Mantalaris for managing and organizing such a great network of courses and secondments, and for making us feel like a big family.

A special thanks to my friends from Spain, Noemi, Tati, Rocío, Leila, Alex, Javi, Pablito, Armada, Corral, Ismael, Toni, Picos, Paula, Joaquin, Santi, and Lucía, for making me feel as if I had never left when I come back home for vacation. And my friends in Lausanne, Katie, my french teacher Camille, Pascal and Valérie, Javier and all the team of SwissThermic, for all the apéros, barbecues and fun times that make me enjoy life in Switzerland. I would also like to thank Maruja and Jose Manuel for taking me in when I first arrived in Lausanne and for their limitless help and support.

I would like to thank all my family, especially my parents, my dear sister Marta, my grandmother Lili, Sesé, Chupi, Padrino, Conchita and Agustín for their love and constant encouragement. Thank you also to my parents-in-law Loly and Juan for embracing me as a family member. And a special thanks to my mother, Loló, the strongest person I know on earth and an example for me to follow. Thank you, mom, for your continuous support and advice during these years and throughout my life.

And finally, Ian, my life companion. Thank you for taking care of me, for all the unconditional support, for always listening to me, for discovering the world together, for your love, for choosing me.

Lausanne, June 2020

" Life isn't about waiting for the storm to pass.
It's about learning how to dance in the rain "
– Vivian Greene

Abstract

Cancer is a leading cause of death in the world, and the mechanisms that underlie this disease are still not completely understood. As cancer develops and progresses, cells undergo a diversity of mutations that sustain their rapid proliferation and the evasion of the immune system. Cancer cells alter their configuration and organization, exhibiting abnormal phenotypes and changes in functionality. The complexity of cancer lies in their heterogeneity and variability among patients, which challenges the current therapies and drug targets. In the last decades, ten hallmarks of cancer cells have been recognized, including alterations in metabolism and the signaling pathways.

The sequencing of the human genome and the advances in *omics* data processing allowed to generate metabolic and signaling networks for human cells at a genome-scale, enlightening the detailed biochemistry and signal transduction processes occurring in human cells, and enabling to study human metabolism and signaling pathways at a systems level. However, the complexity of these networks hinders a consistent and concise physiological representation. In the field of systems biology, mathematical models and computational methods are derived to describe cellular processes based on experimental data and the biological networks. Furthermore, these models have proven to be valuable in understanding the genotype-phenotype relationship of cells and to formulate new hypotheses to guide experimental design.

In this thesis, we present modeling approaches and computational methods to investigate the metabolic and signaling alterations in cancer cells and overcome the challenges arising from biological networks of such size and complexity. Firstly, we curated the thermodynamic properties for all the compounds and reactions in the human metabolic genome-scale models (GEMs) Recon 2 and Recon 3D to guarantee the consistency of the predictions with the bioenergetics of the cell. Moreover, we developed a workflow (redHUMAN) for reconstructing reduced-size models that focus on parts of metabolism relevant to a specific physiology, and we introduce a novel method to

account for the cellular interactions with the extracellular medium. Using redHUMAN, we reduced the human GEMs around pathways that are altered in cancer physiology. Secondly, we applied a set of computational methods to integrate *omics* data into the reduced version of Recon 3D to build metabolic models for breast, colon, and ovarian cancers. These models were used to study how different cancer cells use the metabolic pathways to function and survive and how the underlying genetic deregulations affect the metabolic tasks. Thirdly, we developed a method (CONSIGN) to contextualize signaling networks to a specific type of cell under particular conditions, maximizing the consistency with experimental data. We used this method to generate a breast cancer-specific signaling network for the transcription factor MYC. Finally, we created an integrated model of signaling and metabolic models by accounting for the regulation of metabolic genes by transcription factors. We analyzed the interactions of the MYC signaling network in the breast cancer metabolic model.

The work in this thesis demonstrates the potential of metabolic and signaling models to identify and infer the genetic origins and the microenvironment effects in the transformed phenotype of cancer cells, marking a step forward towards the study of drug targets and biomarkers.

Keywords

human cells, cancer, metabolism, signaling, mathematical models, context-specific models, omics data integration, enrichment analysis, pathway deregulation, integrated signaling-metabolic models.

Résumé

Le cancer est l'une des principales causes de décès dans le monde et les mécanismes qui sous-tendent cette maladie ne sont pas encore complètement compris. À mesure que le cancer se développe et progresse, les cellules subissent une diversité de mutations qui soutiennent leur prolifération rapide et l'évasion du système immunitaire. Les cellules cancéreuses modifient leur configuration et leur organisation, présentant des phénotypes anormaux et des changements de fonctionnalité. La complexité du cancer réside dans leur hétérogénéité et leur variabilité entre les patients, ce qui remet en question les thérapies et les cibles médicamenteuses actuelles. Au cours des dernières décennies, dix caractéristiques des cellules cancéreuses ont été reconnues, y compris des altérations du métabolisme et des voies de signalisation.

Le séquençage du génome humain et les progrès du traitement des données omiques ont permis de générer des réseaux métaboliques et de signalisation pour les cellules humaines à l'échelle du génome, éclairant les processus détaillés de biochimie et de transduction des signaux que se produisant dans les cellules humaines, et permettant d'étudier le métabolisme humain et la signalisation au niveau des systèmes. Cependant, la complexité de ces réseaux entrave une représentation physiologique cohérente et concise. Dans le domaine de la biologie des systèmes, des modèles mathématiques et des méthodes de calcul sont dérivés pour décrire les processus cellulaires basés sur des données expérimentales et les réseaux biologiques. De plus, ces modèles se sont révélés utiles pour comprendre la relation génotype-phénotype des cellules et formuler de nouvelles hypothèses pour guider la conception expérimentale.

Dans cette thèse, nous présentons des approches de modélisation et des méthodes de calcul pour étudier les altérations métaboliques et de signalisation dans les cellules cancéreuses et surmonter les défis découlant de réseaux biologiques d'une telle taille et complexité. Premièrement, nous avons identifié les propriétés thermodynamiques de tous les composés et réactions dans les modèles métaboliques humains à l'échelle du

génome (GEM) Recon 2 et Recon 3D pour garantir la cohérence des prédictions avec la bioénergétique de la cellule. De plus, nous avons développé un méthode (redHUMAN) pour reconstruire des modèles de taille réduite qui se concentrent sur des parties du métabolisme pertinentes pour une physiologie spécifique, et nous introduisons une nouvelle méthode pour tenir compte des interactions cellulaires avec le milieu extracellulaire. En utilisant redHUMAN, nous avons réduit les GEM humains autour des voies qui sont modifiées dans la physiologie du cancer. Deuxièmement, nous avons appliqué un ensemble de méthodes de calcul pour intégrer les données omiques dans la version réduite de Recon 3D pour construire des modèles métaboliques pour les cancers du sein, du côlon et des ovaires. Ces modèles ont été utilisés pour étudier comment différentes cellules cancéreuses utilisent les voies métaboliques pour fonctionner et survivre et comment les dérégulations génétiques sous-jacentes affectent les tâches métaboliques. Troisièmement, nous avons développé une méthode (CONSIGN) pour contextualiser les réseaux de signalisation à un type spécifique de cellule dans des conditions particulières, en maximisant la cohérence avec les données expérimentales. Nous avons utilisé cette méthode pour générer un réseau de signalisation spécifique au cancer du sein pour le facteur de transcription MYC. Enfin, nous avons créé un modèle intégré des modèles de signalisation et métaboliques en tenant compte de la régulation des gènes métaboliques par les facteurs de transcription. Nous avons analysé les interactions du réseau de signalisation MYC dans le modèle métabolique du cancer du sein.

Les travaux de cette thèse démontrent le potentiel des modèles métaboliques et de signalisation pour identifier et inférer les origines génétiques et les effets du microenvironnement dans le phénotype transformé des cellules cancéreuses, marquant une étape vers l'étude des cibles de médicaments et des biomarqueurs.

Mots-clés

cellules humaines, cancer, métabolisme, signalisation, modèles mathématiques, modèles spécifiques au contexte, intégration des données omiques, analyse d'enrichissement, dérégulation des voies, modèles métaboliques de signalisation intégrés

Contents

Acknowledgements ..... **v**

Abstract ..... **ix**

Keywords ..... **x**

Résumé ..... **xi**

Mots-clés ..... **xii**

List of Figures ..... **xvii**

List of Tables ..... **xxiii**

List of Abbreviations ..... **xxv**

Chapter 1 Introduction **1**

1.1 The biology of cells 1

1.2 Cancer, a heterogeneous disease 2

1.3 Mathematical models of biological networks 4

1.3.1 Modeling cellular metabolism 4

1.3.2 Modeling signal transduction pathways 6

1.3.3 Modeling integrated biological networks 7

1.3.4 Methods in Systems Biology 7

1.4 Motivation and objectives of this thesis 10

1.5 Structure of this thesis 12

References 13

Chapter 2 Analysis of human metabolism and growth media using reduced thermodynamically curated genome-scale models. **19**

2.1 Introduction 19
2.2 Results 21
2.2.1 Overall workflow 21
2.2.2 Thermodynamic curation of the human GEMs (Step 1) 23
2.2.3 Subsystem selection to build the core (Step 2) 23
2.2.4 Network expansion (Step 3) 24
2.2.5 Extracellular medium connection (Step 4) 25
2.2.6 Biosynthetic reactions generation (Step 5) 28
2.2.7 Data integration and metabolic tasks (Step 6) 32
2.2.8 Physiology analysis 36
2.3 Discussion 37
2.4 Materials and Methods 40
2.4.1 Experimental data for leukemia cell lines 40
2.4.2 Thermodynamic curation of the genome-scale models (GEMs) 40
2.4.3 TFA: thermodynamics-based flux analysis 41
2.4.4 iMM: characterizing the extracellular *in silico* minimal media 42
2.4.5 redGEM, redGEMX, and lumpGEM: reducing human GEMs 43
2.4.6 Software 45
2.4.7 Data and code availability 46
2.5 Author contribution 46
References 47
Appendix A 54
Chapter 3 Model-based data integration and minimal network enrichment to identify metabolic differences across cancer types 65
3.1 Introduction 65
3.2 Results 67

3.2.1 Workflow overview 67
3.2.2 Building cancer-specific models 69
3.2.3 Cancer phenotype analysis with metabolic models 78
3.2.4 From pathways to minimal networks 82
3.2.5 Minimal network deregulation and enrichment analysis 85
3.3 Discussion 87
3.4 Materials and Methods 89
3.4.1 Exometabolomics, transcriptomics data and gene expression per cell line 89
3.4.2 Cancer-type specific data 89
3.4.3 Human generic metabolic model 90
3.4.4 Integrating context-specific metabolomics and fluxomics data 90
3.4.5 Defining context-specific transport reactions 90
3.4.6 Thermodynamic-flux variability analysis 92
3.4.7 Integration of context-specific expression data 93
3.4.8 Gene and enzyme essentiality analysis 95
3.4.9 Formulating metabolic tasks for cancer physiology 95
3.4.10 Generating Minimal Networks for the metabolic tasks 96
3.4.11 Minimal Network Enrichment Analysis (MiNEA) 97
3.5 Author contribution 98
References 99
Appendix B 106
Chapter 4 Integrating signaling and metabolic pathways to analyze the function of the transcription factor MYC in breast cancer 113
4.1 Introduction 113
4.2 Results 116

4.2.1 Method overview 116
4.2.2 The transcription factor MYC and its relation to cancer 117
4.2.3 A small signaling pathway for MYC 118
4.2.4 Interactions of the signaling pathway and metabolism 123
4.2.5 Upstream signaling pathway for MYC and metabolism of breast cancer 126
4.3 Discussion 134
4.4 Materials and Methods 136
4.4.1 Gene and protein expression data 136
4.4.2 Human breast cancer metabolic model 136
4.4.3 Mapping transcription factors to metabolism 137
4.4.4 Reconstruction of signaling pathways from REACTOME 137
4.4.5 Discrete formulation of rules for GPRs 137
4.4.6 Discrete formulation of rules for signaling interactions 139
4.4.7 Discrete formulation of rules for the regulation of genes by transcription factors 140
4.4.8 Contextualization of signaling networks (CONSIGN) method 140
4.5 Author contribution 142
References 143
Appendix C 149
Chapter 5 Conclusions 153
5.1 Conclusions 153
5.2 Future perspectives 157
References 160
Curriculum Vitae 163

List of Figures

Figure 1.1. **The biological hallmarks of cancer.** A total of ten hallmarks of cancer have been recognized. In this thesis, we focus on the study of two of them, namely the deregulating cellular energetics and sustaining proliferative signaling. The first of them involves alterations at the metabolic level, and the second one includes malfunctions at the signal transduction pathways. Figure adapted from [12]. 3

Figure 1.2. **Size of the human GEMs.** The integration of new information in the GEMs has expanded their size and complexity over the years. The two GEMs highlighted in the graph (magenta) are the basis for the models used in the work presented in this thesis..... 5

Figure 1.3. **Heterogeneity in the tumor microenvironment**. A variety of cells cohabit within the tumor microenvironment, including the cancer cells, which exhibit heterogeneous phenotypes among them, given by the set of mutations that they experience and by the different access to nutrients, immune cells, and healthy cell in the surrounding tissue. Figure adapted from [56]. 10

Figure 2.1. **redHUMAN: workflow to systematically reconstruct thermodynamic-curated reduced models.** (1) Thermodynamic curation: the Gibbs free energy of compounds and reactions are estimated and used to define the reaction directionality. (2) Subsystem selection: the subsystems relevant for the study are selected. (3) Network expansion: the initial subsystems are connected using reactions from the GEM to generate a core network. (4) Extracellular medium connection: the pathways that connect the extracellular medium components to the core network are identified. (5) Biosynthetic reaction generation: the pathways required to produce the biomass building blocks are classified. (6) Data integration and consistency checks: experimental values are integrated and the model is verified through consistency checks. 22

Figure 2.2. **Thermodynamic curation of human GEMs**. (A) Thermodynamics for the unique compounds in Recon 2 (orange) and Recon 3D (blue). The percentage is relative to the total number of unique compounds. (B) Size of the core network when the expansion is performed for different degrees. (C)

Number of reactions that pairwise connect the subsystems for Recon 2 (values below the diagonal) and Recon 3D (values above the diagonal) for degree D = 1..... 25

Figure 2.3. **Extracellular medium utilization** (A) Extracellular medium composition defined in the models. (B) Graph of the subnetwork from Recon 2 for the uptake of L-histidine and the medium components required for its metabolism. Green represents the metabolites from the subnetwork, and orange represents the metabolites of the core network where the subnetwork is connected. In blue, the medium metabolite under study (L-histidine) and in pink, the extracellular metabolites co-utilized to metabolize L-histidine. The pathway starts with the transport of L-histidine from the extracellular space to the cytosol, where it is sequentially transformed into urocanate (urcan_c), 4-imidazolone-5-propanoate (4izp_c), N-formimidoyl-L-glutamate (forglu_c), L-glutamate (glu_L), 5-formiminotetrahydrofolate (5forthf_c), 5-10-methenyltetrahydrofolate (methf_c), and 10-formyltetrahydrofolate (10thf_c). 4-Aminobutanoate (4abut_c) is converted to L-glutamate through a reaction from the subsystem glutamate metabolism, and finally, L-glutamate is connected to the TCA cycle. 26

Figure 2.4. **Biosynthesis of biomass building blocks.** (A) Size of lumped reactions for Recon 2 and Recon 3D, and the corresponding number of alternatives to synthesize the BBBs that cannot be produced by the core nor uptaken from the extracellular medium. (B-C) Subnetwork for the synthesis of phosphatidylserine. Orange represents the metabolites from the core network. Blue represents the metabolites from the subnetwork for phosphatidylserine synthesis. Pink represents the extracellular metabolites. Phosphatidylserine synthesis starts from the core metabolites glycerol 3-phosphate, from glycolysis, and acetyl CoA, from TCA. In a first reaction, acetyl CoA is transformed into malonyl CoA. The next reaction (KAS8) represents the synthesis of palmitate in the elongation cycle [54]. A CoA molecule is attached to palmitate to form palmitoyl CoA, from which the two generic fatty acids are derived. These two generic fatty acids are attached to glycerol 3-phosphate to form lysophosphatidic acid and phosphatidic acid. Finally, serine is attached to phosphatidic acid to form phosphatidylserine (B) Subnetwork from Recon 2 and corresponding lumped reaction. (C) The four alternative subnetworks of minimum size from Recon 3D. Phosphatidic acid can be produced with two generic fatty acids or with one generic fatty acid and the essential fatty acid linoleic acid (light blue reactions). Phosphatidylserine can be directly produced from phosphatidic acid by attaching serine (green reaction) or through the

formation of phosphatidylcholine (red reaction) and then changing choline for serine (orange reaction). 29

Figure 2.5. **Model validation through metabolic tasks and consistency checks**. (A) The 57 metabolic tasks tested in the generated reduced models. R2, R3: Recon 2, Recon 3D reduced model with one lumped reaction per BBB. R2s, R3s: Recon 2, Recon 3D reduced model with Smin. Classification of metabolic tasks in those captured by the models (green) and those not captured by the models (red). MT1: rephosphorylation of nucleoside triphosphates, MT2: de novo synthesis of nucleotides, MT3: uptake of essential amino acids, MT4: de novo synthesis of key intermediates, MT5: de novo synthesis of other compounds, MT6: protein turnover, MT7: electron transport chain and TCA, MT8: beta oxidation of fatty acids, MT9: de novo synthesis of phospholipids, MT10: vitamins and co-factors, MT11: growth. (B) Gene essentiality of the reduced models and their corresponding GEM. R2s has 829 genes associated to reactions, 37 of which are essential both in the reduced model and in Recon 2 and 12 are essential only in the reduced model. R3s has 828 genes associated to reactions, from which 23 are essential in both the reduced model and Recon 3D. The reduced model presents an additional 44 essential genes. (C) Thermo-flux variability analysis (TVA) for reactions in the reduced models. Orange represents fluxes in the reduced Recon 2 model and blue represents fluxes in the reduced Recon 3D models. The black lines correspond to the fluxes in the GEM. 35

Figure 2.6. **redGEMX method**. (A) Classification of the reactions from the GEM into core (green) and non-core reactions (orange), and classification of the extracellular metabolites from the GEM into those that are part of the medium that we want to connect (blue), those that are present in the core (pink), and the others (grey). The algorithm will block the non-core reactions that involve only extracellular metabolites as well as the boundary and transport reactions of the metabolites that are not part of the medium (grey). (B) The algorithm finds the minimal set of reactions that are required to connect each of the medium metabolites (blue) to the core network, uses the core network to balance the reactions, and secretes metabolites from the medium (blue or pink). 45

Figure 3.1. **Overview of the workflow for data integration and pathway enrichment analysis.** The workflow is divided in two parts. In the first part, *Data integration,* physiology-specific models are generated. The network topology and the network physiology are defined by integrating transcriptomics, proteomics, metabolomics, and fluxomics data into the generic metabolic

model. In the second part, *Pathway deregulation,* the physiology specific models are used to generate minimal networks that represent a set of metabolic tasks for the study of specific cancer phenotypes. Then, the method MiNEA is used to perform minimal network enrichment analysis for the metabolic tasks. 69

Figure 3.2. **Analysis of cancer-specific transport reactions.** (A) Expression profile for the enzymes of 1026 transport reactions across the three cancer types, breast cancer, colon cancer, and ovarian cancer. Green represents upregulated enzymes, red downregulated enzymes, and white those that are not deregulated. (B) Transport reactions kept in the cancer models based on gene expression and the model requirements for growth. Overlap of transport reactions per cancer type based on their deregulation and the model requirements (U: upregulated, N: not deregulated, D: downregulated, M: required by the model). 73

Figure 3.3. **Deregulation of subsystems in the context-specific models**. (A) Comparison of deregulated subsystems considering all the transcriptomics data and considering only the deregulated reactions that can have a metabolic rate consistent with the transcriptomics data and are common in all alternatives. (B) Percentage of deregulated subsystems based on the common reactions across alternatives whose reaction rates can be consistent with the transcriptomics data for each cancer type. 77

Figure 3.4. **Cancer phenotype analysis**. (A) Cancer-specific models. Number of metabolites, reactions, genes, and subsystems in each cancer-specific model. (B) Reaction rate variability analysis for the three cancer types. 79

Figure 3.5. **Essentiality analysis**. (A) Gene essentiality for the three cancer types performed, taking into account omics data, thermodynamics, and expression constraints. (B) Enzyme essentiality analysis for the three cancer types with the data imposed in the network for thermodynamics, metabolite concentrations, reaction rates, and consistency with transcriptomics data. 81

Figure 3.6. **Minimal Networks**. Minimal networks vs classical pathway. (A) representation of the classical pathway in KEGG for the synthesis of phosphatidyl-serine and the corresponding minimal network, which also includes the upstream pathways, in this case, glycolysis and TCA. (B) Minimal networks representing the phenotypes of the Warburg effect for breast, colon, and ovarian cancers. For the sake of simplicity in the visualization the networks do not include all the reactions from the MiN, but the central pathways that show the main differences among them. 85

Figure 3.7. **Deregulation of metabolic tasks**. Most significantly deregulated ($p<0.001$) tasks for each cancer type. 86

Figure 4.1. **Method overview.** Signaling events are modeled using Boolean rules that represent protein activation, complex formation, reaction activation, and reaction inhibition. Upon binding of the ligand and receptor (L:R), the signaling proteins activate (P_i) transmitting the signal. When a protein or a complex (P_5:P_6) activates a transcription factor (TF_1), we assume that the corresponding regulated gene (G_1) is active. We use Boolean rules to model the regulation of the expression of a gene by a transcription factor. Then, the active gene transcribes to the corresponding enzyme (E_i), which is produced to an adequate degree to generate flux through the corresponding reaction (v_i). The gene-enzyme-reaction relationship is modeled by Boolean rules following the GPR rules described in the metabolic model. Finally, the reactions follow the stoichiometric and thermodynamic constraints defined in the constraint-based formulation of the metabolic model. 117

Figure 4.2. **MYC signaling pathway**. (A) A branch of the MYC signaling network upstream ten reactions from MYC. Proteins and Genes are considered in the workflow as observable states. We integrate data for the observable states if they are part of the dataset. (B) Two alternative states of the network consistent with the data obtained as solutions of the MILP to maximize consistency after data integration. 121

Figure 4.3. **MYC subnetwork connected to the metabolic network**. MYC is a transcription factor that regulates 12 metabolic genes present in Recon 3D. Specifically, it promotes the expression of argininosuccinate synthase (*ASS1*), ADP-ribosyl cyclase (*CD38*), Galactoside 3(4)-L-fucosyltransferase (*FUT3*), L-lactate dehydrogenase A chain (*LDHA*), Ornithine decarboxylase (*ODC1*), Thioredoxin-dependent peroxide reductase (*PRDX3*), CMP-N-acetylneuraminate-beta-galactosamide-alpha-2,3-sialyltransferase 1 (*ST3GAL1*), CMP-N-acetylneuraminate-beta-1,4-galactoside alpha-2,3-sialyltransferase (*ST3GAL3*), CMP-N-acetylneuraminate-beta-galactosamide-alpha-2,3-sialyltransferase 4 (*ST3GAL4*), and it inhibits the expression of Branched-chain-amino-acid aminotransferase (*BCAT1*), Choline kinase alpha (*CHKA*), and Proline dehydrogenase 1 (*PRODH*). Five of those genes encode enzymes that catalyze reactions that are part of the breast cancer-specific metabolic model used in this study, namely, L-lactate dehydrogenase (LDH_L), Glyoxylate oxidase (GLXO1), Ornithine Decarboxylase (ORNDC and

HMR_4422), Choline kinase (CHOLK) and Proline dehydrogenase (PROD2m). 124

Figure 4.4. **Distribution of the network species that are constrained to be active or inactive based on the transcriptomics data and the signaling network constraints.** Genome-wide classification of the Reactome pathways related to active (A) and inactive (B) species whose states where integrated and inferred from the consistency between the transcriptomics data and the signaling network. The graph was obtained with Reactome Data Analysis Tool. The color code denotes the over-representation of the pathway based on the p-value. Light grey indicates pathways that are not significantly over-represented. 130

Figure 4.5. **Distribution of the network species that are constrained to be active or inactive based on the transcriptomics data and the integrated network constraints.** Genome-wide classification of the Reactome pathways related to active (A) and inactive (B) species whose states where integrated and inferred from the consistency between the transcriptomics data and the integrated network. The graph was obtained with Reactome Data Analysis Tool. The color code denotes the over-representation of the pathway based on the p-value. Light grey indicates pathways that are not significantly over-represented. 133

Figure 5.1. **Simulating the heterogeneity in the tumor microenvironment.** We have now the models and methods to simulate the behavior of the different type of cells that populate the tumor microenvironment allowing to investigate the similarities and differences among them. Figure adapted from [9]. 157

List of Tables

Table 1.1. Methods developed in Systems Biology for the study of metabolism and signalling pathways in the cells. Highlighted in bold are those used in this thesis, highlighted in bold and blue are those derived in the work of this thesis..... 8

Table 2.1**. Statistics on the generated reduced metabolic models**. The models were generated from the human GEMs Recon 2 and Recon 3D. For each GEM, two reductions were performed considering either one lumped reaction per BBB (one per BBB) or all the alternatives lumped reaction with subnetworks of minimum size (Smin). 32

Table 3.1. **MiNs per cancer-type for different phenotypes**. MiN: Minimal network, HFRs: High-frequency reactions. Phenotypes and metabolic tasks associated: Warburg effect (lactate), glutamine addiction (glutamate), stress response (superoxide anion and hydrogen peroxide), energy metabolism (ATP through the electron transport chain), serine pathway (serine and glycine), reprogramming of the pentose phosphate pathway (ribose-5P) and phospholipid synthesis (phosphatidyl-serine, phosphatidyl-choline, and phosphatidyl-inositol). 84

Table 4.1. **Alternative states of the negative feedback loop network for MYC**. Seven alternative states of the network that are consistent with the three species active in the transcriptomics data (blue). The state of another three species can be inferred from the integration of the transcriptomics data (orange). 122

Table 4.2. **Transcription factors and metabolic genes**. Transcription factors (TFs) that regulate metabolic genes present in the breast cancer specific model. The corresponding regulation is specified as activation [A] or repression [I]. The two genes used as examples in the text are highlighted in blue and green in the tables; 127

Table 4.3. **Integration of data and consistency analysis**. Comparison of the maximum consistency and the number of alternatives 131

List of Abbreviations

ACP	Acyl carrier protein
ACSN	Atlas of Cancer Signaling Network
BBB	Biomass Building Block
CCLE	Cancer Cell Line Encyclopedia
CHEBI	Chemical Entities of Biological Interest
CoA	Coenzyme A
CONSIGN	Method: Contextualization of signaling networks
DNA	Deoxyribonucleic acid
ETFL	Method: Expression and Thermodynamics Flux models
FASTCORE	Method: Fast reconstruction of core networks
FBA	Flux Balance Analysis
FVA	Flux Variability Analysis
GCM	Group Contribution Method
gDW	gram dry weight
GECKO	GEM with Enzyme Constraints using Kinetic and Omics data
GEM	Genome-scale metabolic model
GIMME	Method: Gene Inactivity Moderated by Metabolism and Expression
GPR	Gene-Protein-Reaction rule
h	hour
HFR	High Frequency Reaction
HMDB	Human Metabolome Database
iDREAM	Method: Integrated Deduced and Metabolism
iFBA	Method: Integrated Flux Balance Analysis
iMM	Method: *in silico* Minimal Media
IOMA	Method: Integrative Omics-Metabolic Analysis
KEGG	Kyoto Encyclopedia of Genes and Genomes
lumpGEM	Method: lumped reactions subnetworks generation
M	molar
MBA	Method: Model-Building Algorithm
mCADRE	Method: metabolic Context-specificity Assessed by Deterministic Reaction Evaluation
ME-model	Metabolism and expression model
MILP	Mixed-Integer Linear Programming

MiN	Minimal network
MINEA	Method: minimal network enrichment analysis
mmol	milimol
mRNA	Messenger ribonucleic acid
MT	Metabolic task
MYC	BHLH Transcription Factor
MYC	MYC Proto-Oncogene
NCI60	National Cancer Institute cell line panel for 60 cancer cell lines
ODE	Ordinary Differential Equation
ORACLE	Method: Optimization and Risk Analysis of Complex Living Entities
PPP	Pentose Phosphate Pathway
PROM	Method: Probabilistic Regulation Of Metabolism
redGEM	Method: reduction of genome-scale metabolic models
redGEMX	Method: connection of extracellular metabolites to core
redHUMAN	Method: reduce the size of the human metabolic models
rFBA	Method: Regulatory Flux Balance Analysis
RNA	Ribonucleic acid
ROS	Reactive Oxygen Species
s.t.	subject to
SEED	Identificator number from modelSEED database
SMILE	Simplified Molecular Input Line Entry Specification
SR-FBA	Method: State Regulatory Flux Balance Analysis
TCA cycle	Tricarboxylic acid cycle
TCGA	The Cancer Genome Atlas Program
TEX-FBA	Method: Thermodynamics and expression flux balance analysis
TFA	Method: Thermodynaics-based Flux Balance Analysis
TIGER	Method: Integrative Network Inference for Tissues
tINIT	Method: task-driven integrative network inference for tissues
TRFBA	Method: Transcriptional Regulated Flux Balance Analysis
TVA	Thermodynamic Flux Variability Analysis
uFBA	unsteady-state Flux Balance Analysis

Chapter 1 Introduction

In this chapter, we introduce the background for the work performed in this thesis. We present the biology of cells, the genotype-phenotype relationship, and the alterations that emerge in damaged cells. In particular, we focus on cancer and the hallmarks defined for this disease in the last years. We review the current modeling approaches available for the study of metabolic networks and signal transduction networks. Finally, we present the motivation and objectives, and the structure of this thesis, including a brief description of each chapter.

1.1 The biology of cells

Cells are considered the fundamental units of life. They organize to form tissues, organs, and complex organisms. Cells develop their structure according to their function, and they do it in a process called differentiation, where each type of cell expresses a specific set of genes that give to the cell its behavior, metabolism, and physiology [1]. Gene expression is coordinated by transcription factors to display tissue-specific phenotypes. Transcription factors enhance or repress the expression of the genes by promoting the process of copying the DNA sequence for specific genes into RNA molecules, which are then translated by the ribosomes into amino acid sequences. The resulting macromolecule folds into a unique three-dimensional configuration and becomes a functional protein. Post-translational modifications give the final form to the protein that specializes in a function, such as signal transduction, structural component, or enzyme. Enzymes are the proteins responsible for catalyzing metabolic reactions that define the

biochemistry inside the cell [2]. The collection of functional proteins and the physiological conditions surrounding the cell shape the phenotype of the cells.

Throughout the life of a cell, what is known as cell cycle, the functions and processes occurring inside the cell, including replication, DNA repair, growth and metabolism, protein synthesis, and motility, are tightly regulated by a group of proteins. Damaged cells that cannot correctly perform these processes, due to errors, follow precise instructions to undergo cell cycle arrest or apoptosis [3]. Cellular damage can be of different types and may occur at various levels. For example, viruses deregulate the host cell cycle to promote viral replication, and sometimes they interfere with the immune cells propagating the infection [4, 5], in Alzheimer's disease, neurons reactivate the cell cycle resulting in apoptosis and damaging the brain tissue [6, 7], and cancer cells contain mutations in several genes that encode cell cycle regulators, promoting uncontrolled cell division [8, 9]. In some cases, the diseased cells develop mechanisms to avoid apoptosis. As a result, they survive and allow the disease to evolve further.

Understanding the origin of the alterations that arise in diseased cells will help to find better targets and to create more effective therapies. The work in this thesis centers on the study of the alterations that occur in cancer cells, and it identifies the underlying mechanisms that contribute to the phenotypic similarities and differences across different types of cancer.

1.2 Cancer, a heterogeneous disease

Cancer is a major cause of death worldwide; its complexity and variability among patients makes personalized treatment challenging and difficult. Cancer cells undergo a diversity of genetic mutations, simultaneously deregulating a variety of cellular processes, including metabolism, and cell growth and division. As they develop, cancer cells promote angiogenesis to provide the tumor cells with a source of oxygen and nutrients, they fail to undergo apoptosis upon DNA damage, and they experience metabolic changes that support their increased proliferation rate.

Furthermore, the tumor microenvironment is shaped by the tumor cells to allow growth and proliferation. Cancer cells manipulate the surrounding cells by depleting essential nutrients and accumulating immunosuppressive metabolites [10-12]. In the last decade,

Hanahan and Weinberg have published ten distinct hallmarks of cancer, consisting of, proliferative signaling, evading growth suppressors, avoiding immune destruction, enabling replicative immortality, promoting inflammation, activating invasion and metastasis, inducing angiogenesis, enabling genome instability and mutation, resisting cell death, and deregulating cellular energetics and metabolism (Figure 1.1). The combination of these alterations causes the accumulation of different phenotypes that populate the heterogeneity in the tumor microenvironment [12].

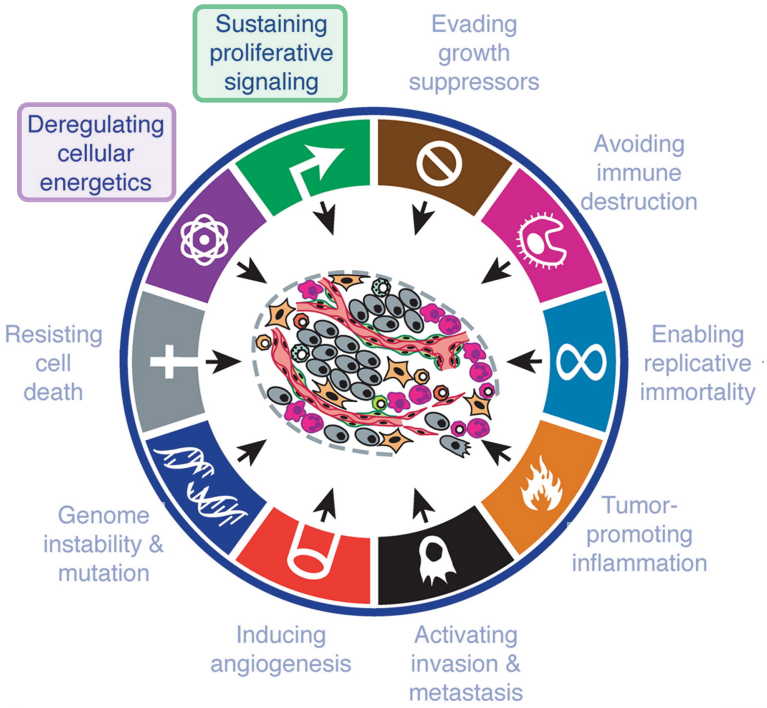


Figure 1.1. **The biological hallmarks of cancer.** A total of ten hallmarks of cancer have been recognized. In this thesis, we focus on the study of two of them, namely the deregulating cellular energetics and sustaining proliferative signaling. The first of them involves alterations at the metabolic level, and the second one includes malfunctions at the signal transduction pathways. Figure adapted from [12].

The heterogeneity of cancer cells caused by genetic modifications and their diverse access to nutrients interferes with the discovery of significant targets for treatments. Moreover, cancer cells have developed mechanisms to adapt to the therapies and defeat the drugs by modifying their genetic activity and rewiring their metabolism, producing drug-resistant surviving cells [13, 14].

Therefore, it is critical to decipher the mutations that cancer cells manifest with respect to their healthy counterparts, the alternative mechanisms that cancer cells have to defend against the treatments, and the orchestrated alterations that allow cancer cells to survive in a strictly-controlled system despite being damaged cells.

1.3 Mathematical models of biological networks

Advances in high-throughput technologies have allowed to analyze tumors at multiple levels, highlighting the overwhelming complexity of the disease and recognizing cancer as a Systems Biology disease [14, 15]. Systems Biology approaches analyze the large amounts of high-quality data collected in the last decades, to gain knowledge about the genotype-phenotype relationship in cancer at a systems level. To this end, mathematical models and computational methods are created to represent an overview of the complexity of cells and their responses to changes in their environment. The models allow for a rigorous study of the complex network of pathways involving gene regulation, signaling, and cell metabolism, and their alterations caused by the genetic mutations occurring in cancer cells [16]. Moreover, these models are used to understand the interactions within the biological networks and to generate and test hypotheses that could help to identify new drug targets and develop better therapies [17].

In the following, we present some of the current Systems Biology approaches applied in the investigation of metabolic and signaling pathways in cells, as well as to study the complex interplay between metabolism and signaling, which is responsible for cell physiology and cell behavior [18].

1.3.1 Modeling cellular metabolism

Metabolism is the set of biochemical reactions within a cell to transform nutrients into energy and cellular building blocks. In the field of Systems Biology, genome-scale metabolic networks (GEMs) were reconstructed based on the annotation of the genome of the organism [19], and they contain the known biochemical reactions occurring inside the cells of the organism. Moreover, GEMs include the annotation of the gene-protein-reaction associations (GPR rules) that relate genes and enzymes, and they are used to elucidate the cellular genotype-phenotype relationship [20].

With the annotation of the human genome sequences in 2001 and 2004 [21, 22], the scientific community reconstructed in 2007 the first genome-scale models for human metabolism, named, Recon 1 [23] and EHMN [24]. These models were curated and refined over the years, and improved versions of the human GEMs were generated, including, HMR [25], Recon 2 [26], HMR 2.0 [27], Recon 2.2 [28], iHsa [29], and the most recent versions Recon 3D [30] and Human1 [31].

These GEMs are used to model metabolism with two types of approaches: steady-state constraint-based models that involve linear equations, and kinetic models, which include ordinary differential equations. Constraint-based models rely on the stoichiometric relation between reactions and metabolites, overcoming the lack of dynamic or kinetic data, required to perform parameter identification in the kinetic models.

During the years, a phylogeny of constraint-based methods has been created to simulate the metabolic behavior of cells, such as satisfy a specific task, optimize the production of compounds of interest, and to predict cellular phenotypes using GEMs [32]. Some of these methods incorporate constraints at steady-state for mass balance (Flux Balance Analysis [33]), enzyme usage (parsimonious Flux Balance Analysis), and thermodynamic laws (Thermodynamic-based Flux Balance Analysis [34, 35]). Such methods have been used to formulate biological hypotheses and to guide the experiments generating new sets of data that could be integrated into the metabolic models to further improve their predictive capabilities. GEMs and the methods developed are powerful platforms for the integration of *omics* data, including transcriptomics, proteomics, metabolomics, and fluxomics.

As our knowledge of human metabolism increases so does the size and complexity of the metabolic models in terms of reactions, metabolites, and genes (Figure 1.2).

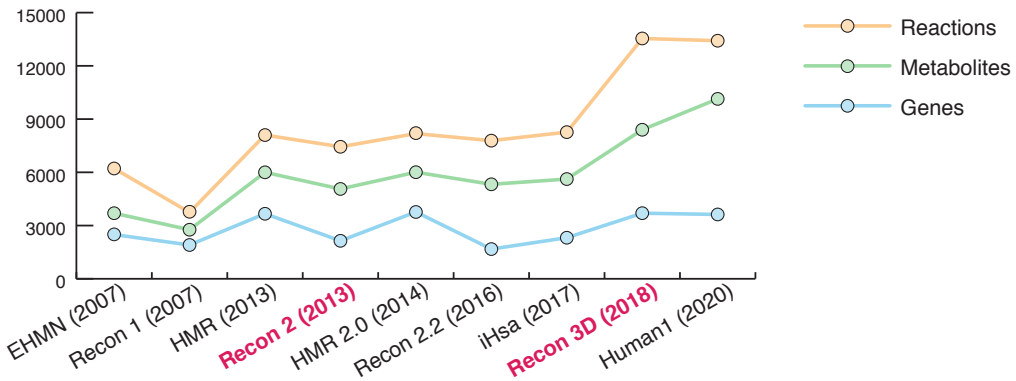


Figure 1.2. **Size of the human GEMs.** The integration of new information in the GEMs has expanded their size and complexity over the years. The two GEMs highlighted in the graph (magenta) are the basis for the models used in the work presented in this thesis.

The ever-increasing size of the human metabolic models hinders their utilization for biological studies, as their increased complexity hampers the analysis of results and increases the computational cost. Furthermore, these networks are reconstructed based on the whole genome of human cells, while a specific type of cell expresses only a portion of those genes. To reduce the models to a more manageable size and to represent a particular type of cell, a plethora of methods was developed to derive reduced-size context-specific models from these generic human models. The context-specific models capture the phenotype of a specific type of cell or tissue by reducing the generic GEM to the reactions that are catalyzed by enzymes expressed in the specific tissue. These methods rely on transcriptomics or proteomics data and metabolic tasks to identify the set of reactions that define the context-specific model [36-40]. These cell-type and tissue-specific models have been successfully used to simulate metabolism in disease and healthy cells as well as to identify biomarkers and drug targets [40]. Furthermore, personalized models that integrate omics data from patient's samples have been useful for precision medicine [18, 41-43].

1.3.2 Modeling signal transduction pathways

Signal transduction is the collection of pathways that are used by cells to transmit the signals received by their receptors. It represents a signal-flow of information in the cell and allows cells to communicate and to respond to external stimuli. Mathematical models and computational simulations help to determine the integrated functions of signaling networks, leading to exciting biological discoveries [44].

Signaling networks can be studied using quantitative kinetic models or qualitative modeling [45]. The quantitative approach uses a continuous system of ordinary differential equations requiring kinetic parameters and elucidating complex mechanisms with high accuracy. The qualitative approach uses a discrete model that identifies which signal transduction pathways lead from an environmental perturbation to a cellular phenotype without providing details on concentrations of components [46]. When modeling small systems focused on specific events, such as a transcription factor binding to a DNA region or a small pathway activating a protein that generates a cellular response, we can have access to specific data that allows to model in detail the kinetics occurring in the system. However, as the size of the system increases, quantitative models become challenging, and therefore the qualitative approach is more suitable for

analysis of large-scale signaling networks and can be used to explore the topological characteristics of the signaling network [46].

The most common qualitative approach is Boolean modeling that represents the states of the proteins in a switch on/off approach representing active or inactive proteins. Logic rules are then used to represent the flow of information through the signaling pathways. Boolean models have been successfully used to model the cellular responses to stimuli or external perturbations and mapped to observed phenotypes [47, 48]. These models are good platforms to integrate *omics* data, they give insight of the overall state of the signaling pathways, they identify regulatory hubs and, they uncover possible targets for pharmacological intervention in diseases [49] [50].

1.3.3 Modeling integrated biological networks

The alterations occurring in cancer cells have been studied independently. However, as a systems disease, there is an important cross-talk between the biological networks that affect the development and progression of the disease. The integration of different biological networks and data-types has been a major challenge in systems biology. Several methods have been derived to integrate metabolic and signaling networks by accounting for the transcriptional regulation of metabolic genes. The first method that attempted to integrate the effects of both networks within Systems Biology was regulatory Flux Balance Analysis [51], which uses a Boolean approach to include the transcriptional regulation of genes. Based on this method, other methods were recently developed, including Probabilistic Regulation Of Metabolism [52], Flex-Flux [54], and Transcriptional Regulated Flux Balance Analysis [55]. These methods integrate in the metabolic model constraints to account for the expression of the genes and their regulatory effects in metabolism, as well as to control the flux through the reactions in the metabolic network. These integrated models allow to analyze the molecular alterations that coordinate at different levels to promote tumor progression.

1.3.4 Methods in Systems Biology

In the following, to improve the readability of this thesis for a broader audience, we summarize some of the currently available Systems Biology methods developed for the study of the metabolic and signaling pathways within the cells (Table 1.1). The list includes the acronym, the long name, and a brief description of each method. We

constrain the list to those mentioned, cited and developed in this thesis, however other methods can be found in the literature.

Table 1.1. Methods developed in Systems Biology for the study of metabolism and signalling pathways in the cells. Highlighted in bold are those used in this thesis, highlighted in bold and blue are those derived in the work of this thesis.

Acronym	Name & Reference	Description
FBA	Flux balance analysis Orth et al. *Nature Biotechnology.* 2010	Constraint-based modelling of metabolic networks imposing stoichiometric constraints and steady-state conditions.
GCM	Group contribution method Jankowski et al. *Biophysical Journal*. 2008	Estimation of the standards Gibb's free energy of formation of compounds in biochemical systems based on molecular substructures.
TFA	Thermodynamics-flux balance analysis Henry et al. *Biophysical Journal*. 2007	Integration of thermodynamic properties of compounds in the model and constraints in the FBA problem. MILP formulation.
INIT	Integrative Network Inference for Tissues Agren et al. *PLoS Comput Biol*. 2012	Generation of tissue-specific models from GEMs by waiting the activation of reactions based on expression. Allows accumulation of metabolites.
tINIT	task-driven integrative network inference for tissues Agren et al. *Mol. Syst. Biol*. 2014	Generation of tissue-specific models with INIT including also constraints to satisfy a set of metabolic tasks specific for the tissue.
GIMME	Gene Inactivity Moderated by Metabolism and Expression Becker et al. *PLoS Comput Biol*. 2008	Reconstruction of context-specific models favouring highly expressed reactions and those related to a metabolic objective.
iMAT	Integrative metabolic analysis tool Zur et al. Bioinformatics. 2010	Constraining the reaction rates in constraint-based models based on the expression of the associated genes. MILP formulation.
mCADRE	metabolic Context-specificity Assessed by Deterministic Reaction Evaluation Wang et al. *BMC Systems Biology.* 2012	Reconstruction of context-specific model imposing weights based on expression data, network structure and metabolic functions.
iMM	In silico minimal medium Tymoshenko et al. *PLoS Comput Biol*. 2015	Classification of the nutrient requirements based on data and the metabolic model constraints. MILP formulation.
MBA	Model-Building Algorithm Jerby et al. *Mol Syst Biol.* 2010	Reconstruction of tissue-specific models from core reactions. It imposes weights on the reactions based on expression data.
redGEM	redGEM Ataman et al. *PLoS Comput Biol*. 2017	Reconstruction of core networks based on a set of initial subsystems. The initial subsystems are expanded and connected based on stoichiometry.
redGEMX	redGEMX Masid et al. *Nat Commun.* 2020	Identification of subnetworks that connect the extracellular components to a core network generated by redGEM.
lumpGEM	lumpGEM Ataman et al. *PLoS Comput Biol*. 2017	Identification of subnetworks required to biosynthesize biomass building blocks from a core network.
redHUMAN	Reduced models for human metabolism Masid et al. *Nat commun*. 2020	Workflow to generate reduced models for human GEMs accounting for the pathways used to metabolize the nutrients and synthesize the BBBs.

uFBA	unsteady-state Flux Balance Analysis Bordbar et al. *Scientific Reports.* 2017	Constraint-based modeling method to integrate time-course metabolomics data and predict metabolic flux states for dynamic systems.
ORACLE	Optimization and Risk Analysis of Complex Living Entities Miskovic et al. *Trends Biotechnol.* 2010	Kinetic models for biological systems derived using constraint-based models as a scaffold and sampling kinetic parameters. Analysis of control coefficients.
GECKO	GEM with Enzyme Constraints using Kinetic and Omics data Sanchez et al. *Mol Syst Biol.* 2017	Formulation that enhances the GEM using enzyme constraints with kinetic and omics data.
ETFL	Expression and Thermodynamics Flux models Salvy et al. *Nat Commun.* 2020	Integration of expression and thermodynamic constraints to constraint-based models to account for enzymes and mRNA levels. MILP formulation.
TEX-FBA	Thermodynamics and expression flux balance analysis Pandey et al. bioRxiv. 2019	Integration of expression and thermodynamic data in constraint-based models by translating gene deregulation into flux rate constraints.
IOMA	Integrative Omics-Metabolic Analysis Yizhak et al. *Bioinformatics.* 2010	Quantitative integration of proteomics and metabolomics data in GEMs. Maximize consistency between measured and kinetically derived fluxes.
FASTCORE	Fast Reconstruction of Compact Context-Specific Metabolic Network Models Vlassis et al. *PLoS Comput Biol.* 2014	Reconstruction of context specific models from an initial set of expressed reactions, minimizing the number of additional reactions for activity of core.
MINEA	Minimal Network Enrichment Analysis Pandey et al. *PLoS Comput Biol.* 2019	Identification of minimal set of reactions required for a metabolic task. Enrichment of the network based on gene expression data.
rFBA	Regulatory Flux Balance Analysis Cover et al. *J. Theor. Biol. 2001*	Integration of boolean rules to account for regulatory events within FBA. Addition of temporary constraints on the metabolic network.
SR-FBA	State Regulatory Flux Balance Analysis Shlomi, et al. *Mol. Syst. Biol.* 2007	Boolean approach to integrate regulatory and metabolic networks by connecting the transcription factors, the metabolic genes and the enzymes.
iFBA	Integrated Flux Balance Analysis Covert, et al. *Bioinformatics*, 2008	Framework to integrate transcriptional and metabolic networks, by combining an FBA model, a Boolean transcriptional network, and a set of ODEs.
PROM	Probabilistic Regulation Of Metabolism Chandrasekaran et al. *PNAS.* 2010	Probabilistic model for the regulatory network based on abundant expression data used to define reaction rates in the metabolic network.
iDREAM	Integrated Deduced and Metabolism Wang, et al. *PLoS Comput Biol.* 2017	Generation of metabolic-regulatory network models using bootstrapping- EGRIN inferred transcriptional factor regulation of genes combined with PROM.
FlexFlux	FlexFlux Marmiesse, et al. *BMC Syst. Biol.* 2015	Integration of regulatory and metabolic networks by bounding the fluxes based on gene expression.
TRFBA	Transcriptional Regulated Flux Balance Analysis Motamedian et al. *Bioinformatics.* 2017	Integration of transcriptional effects in metabolic networks by adding linear constraints to FBA to bound the reaction rates based on gene expression.
TIGER	Toolbox for integrating GEMs, expression and regulation Jensen, et al. *BMC Syst Biol.* 2011	Conversion of Boolean or multilevel rules into a set of mixed integer inequalities and integration of gene expression in GEMs and transcriptional regulation.
CONSIGN	Contextualization of Signaling Networks *Unpublished*, Chapter 3 in this thesis	Integration of gene/protein expression data into signalling networks to maximize the consistency of the network states with the data.

1.4 Motivation and objectives of this thesis

Besides understanding the alterations that emerge at different levels in tumor cells, controlling and interpreting the phenotypic heterogeneity of the tumor microenvironment is crucial to develop successful therapies. Within the tumor environment, there exists an extensive diversity of cells, including the tumor cells, the immune cells, and the healthy cells (Figure 1.3). Systems biology approaches are created to attain an understanding of the cellular modifications occurring in cancer at a systems level. The generated mathematical models and computational methods help to gain insight from the experimental data regarding the tumor cells and to provide context to the data by creating hypotheses about the origins of the alterations and predicting the effects of drugs in the system.

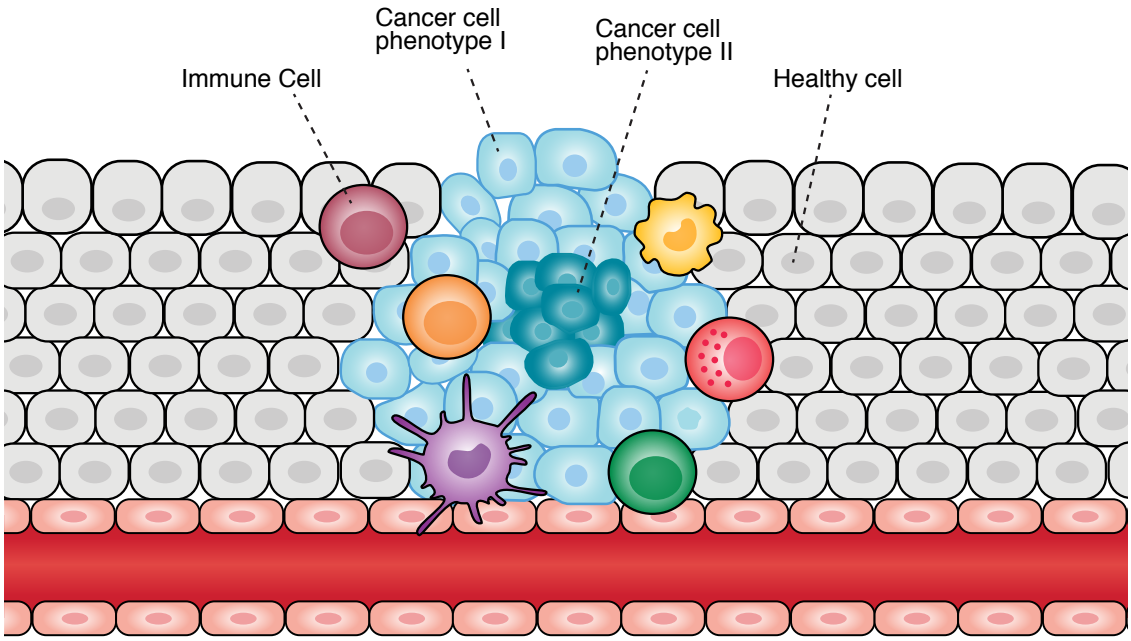


Figure 1.3. **Heterogeneity in the tumor microenvironment**. A variety of cells cohabit within the tumor microenvironment, including the cancer cells, which exhibit heterogeneous phenotypes among them, given by the set of mutations that they experience and by the different access to nutrients, immune cells, and healthy cell in the surrounding tissue. Figure adapted from [56].

In this thesis, we aim to create mathematical models and computational methods that allow to simulate the metabolism and the signaling pathways of human cells and analyze the alterations developed at both levels in tumor cells, as well as, to compare the differences in metabolism and signaling pathways of the different type of cells that populate the tumor microenvironment. Herein, we propose modeling approaches to

overcome some of the current challenges and limitations by further improving the existing algorithms and developing new methods for the study of complex diseases as cancer.

The human GEMs are reconstructed as a collection of the biochemical reactions occurring in any human cell. However, a specific type of cell expresses a portion of the enzymes that catalyze those reactions and lives in a particular environment with access to a set of nutrients. One of the objectives of this thesis was to reconstruct human metabolic models that represented a specific type of cell under particular physiological conditions complying at the same time with the thermodynamic laws that govern the bioenergetic capabilities of the cells. Furthermore, we aimed at developing methods to extract from the GEM a subset of pathways that were of interest for the cellular physiology under study minimizing the loss of information from the other pathways in the network, accounting, up to some degree, for the metabolic pathways required to biosynthesize the cellular building blocks, and keeping consistency with the predictive capabilities of the complete GEM.

The next goal of this thesis was to simulate the metabolism of cancer cells by integrating experimental data into the metabolic model to define the network topology and the network physiology. One of the main questions in biology is the genotype-phenotype relationship, and how we can correlate the deregulation observed in the gene expression profiles to the deregulations seen in the phenotype, such as deregulation in metabolic pathways. Our objective was to develop an approach to assign with a certain degree of confidence deregulation to the metabolic pathways. The cancer-specific developed models were powerful tools to assign deregulation to the metabolic fluxes based on the observed deregulation of the genes that code for the enzymes in the pathways, by testing the capabilities of the network to sustain the observed genotype-phenotype.

Furthermore, an analysis of the essential pathways in the cancer-specific models and the identification of the metabolic pathways that are required to satisfy metabolic functions in the phenotype, allowed us to assign functionality to the deregulations observed.

The final objective of this thesis was to develop a set of methods that permitted the contextualization of signaling networks to a specific physiology and the integration of metabolic networks and signaling networks to study the interplay between signaling and

metabolism. The methodology developed allowed us to interpret the consistency of the observed data and both biological networks.

1.5 Structure of this thesis

In this thesis, we present mathematical and computational methods to navigate the complexity of the human metabolic and signaling networks that allow us to analyze and contextualize the alterations that are present in different cellular stages. In particular, we examine how differences in gene expression across cancer types translate into differences in their metabolism, and how we can simultaneously study signaling and metabolic networks. This thesis is composed of five chapters.

In **Chapter 1**, we introduce the biological and modeling background required for the work performed in this thesis, and we present the motivation and objectives of the thesis. In **Chapter 2,** assign the thermodynamic properties of compounds and reactions in the human GEMs, and we present a novel method (redHUMAN) to generate reduced-sized metabolic models that comprise the pathways relevant for the study. We create a reduced model containing the metabolic pathways that are altered in cancer cells. In **Chapter 3**, we integrate omics data into the reduced model to build cancer-specific metabolic models. We use the cancer-specific models to investigate how different cancer types use different pathways to perform metabolic functions. In **Chapter 4**, we present a novel method (CONSIGN) to generate signaling networks that are consistent with the data. Moreover, we develop an approach to integrate signaling and metabolic networks by including the regulation of metabolic genes by transcription factors. In **Chapter 5**, we summarize the conclusions and future perspectives of the models and methods here presented. Each of the core chapters, i.e., **Chapter 2**, **Chapter 3** and **Chapter 4**, describes a separate manuscript involving, in some occasions, collaborators. At the end of these chapters, we specify the contributions of the thesis author and the collaborators.

References

1. Weinberg, R.A., *The biology of cancer*. 2007, New York: Garland Science.
2. Cummings, M.R., *Human heredity : principles & issues*. 7th ed. 2006, Belmont, Calif.: Thomson Brooks/Cole. xxiii, 457 p.
3. Potten, C.S. and J.W. Wilson, *Apoptosis : the life and death of cells*. 2004, Cambridge ; New York: Cambridge University Press. xvi, 202 p.
4. Bagga, S. and M.J. Bouchard, *Cell Cycle Regulation During Viral Infection. In: Noguchi E., Gadaleta M. (eds) Cell Cycle Control*. Methods in Molecular Biology,. Vol. 1170. 2014, New York: Humana Press.
5. Rouse, B.T. and S. Sehrawat, *Immunity and immunopathology to viruses: what decides the outcome?* Nat Rev Immunol, 2010. **10**(7): p. 514-26.
6. Moh, C., et al., *Cell cycle deregulation in the neurons of Alzheimer's disease.* Results Probl Cell Differ, 2011. **53**: p. 565-76.
7. Herrup, K., *The involvement of cell cycle events in the pathogenesis of Alzheimer's disease.* Alzheimers Res Ther, 2010. **2**(3): p. 13.
8. Vogelstein, B. and K.W. Kinzler, *Cancer genes and the pathways they control.* Nat Med, 2004. **10**(8): p. 789-99.
9. Vermeulen, K., D.R. Van Bockstaele, and Z.N. Berneman, *The cell cycle: a review of regulation, deregulation and therapeutic targets in cancer.* Cell Prolif, 2003. **36**(3): p. 131-49.
10. Bast, R.C., et al., *Holland-Frei cancer medicine*. Ninth ed. 2017, Hoboken, New Jersey: John Wiley & Sons, Inc. ISBN: 978-1-118-93469-2.
11. Hanahan, D. and R.A. Weinberg, *The hallmarks of cancer.* Cell, 2000. **100**(1): p. 57-70.
12. Hanahan, D. and R.A. Weinberg, *Hallmarks of Cancer: The Next Generation.* Cell, 2011. **144**(5): p. 646-674.
13. Housman, G., et al., *Drug resistance in cancer: an overview.* Cancers (Basel), 2014. **6**(3): p. 1769-92.

14. Zecena, H., et al., *Systems biology analysis of mitogen activated protein kinase inhibitor resistance in malignant melanoma.* Bmc Systems Biology, 2018. **12**.

15. Hornberg, J.J., et al., *Cancer: A systems biology disease.* Biosystems, 2006. **83**(2-3): p. 81-90.

16. Laubenbacher, R., et al., *A systems biology view of cancer.* Biochimica Et Biophysica Acta-Reviews on Cancer, 2009. **1796**(2): p. 129-139.

17. Du, W. and O. Elemento, *Cancer systems biology: embracing complexity to develop better anticancer therapeutic strategies.* Oncogene, 2015. **34**(25): p. 3215-3225.

18. Angione, C., *Human Systems Biology and Metabolic Modelling: A Review-From Disease Metabolism to Precision Medicine.* Biomed Res Int, 2019. **2019**: p. 8304260.

19. Thiele, I. and B.O. Palsson, *A protocol for generating a high-quality genome-scale metabolic reconstruction.* Nature Protocols, 2010. **5**(1): p. 93-121.

20. Yizhak, K., et al., *Phenotype-based cell-specific metabolic modeling reveals metabolic liabilities of cancer.* Elife, 2014. **3**.

21. Venter, J.C., et al., *The sequence of the human genome.* Science, 2001. **291**(5507): p. 1304-51.

22. Collins, F.S., et al., *Finishing the euchromatic sequence of the human genome.* Nature, 2004. **431**(7011): p. 931-945.

23. Duarte, N.C., et al., *Global reconstruction of the human metabolic network based on genomic and bibliomic data.* Proceedings of the National Academy of Sciences of the United States of America, 2007. **104**(6): p. 1777-1782.

24. Ma, H., et al., *The Edinburgh human metabolic network reconstruction and its functional analysis.* Mol Syst Biol, 2007. **3**: p. 135.

25. Mardinoglu, A., et al., *Genome-scale metabolic modelling of hepatocytes reveals serine deficiency in patients with non-alcoholic fatty liver disease.* Nature Communications, 2014. **5**.

26. Thiele, I., et al., *A community-driven global reconstruction of human metabolism.* Nature Biotechnology, 2013. **31**(5): p. 419-+.

27. Pornputtapong, N., I. Nookaew, and J. Nielsen, *Human metabolic atlas: an online resource for human metabolism.* Database-the Journal of Biological Databases and Curation, 2015.

28. Swainston, N., et al., *Recon 2.2: from reconstruction to model of human metabolism.* Metabolomics, 2016. **12**(7).

29. Blais, E.M., et al., *Reconciled rat and human metabolic networks for comparative toxicogenomics and biomarker predictions.* Nature Communications, 2017. **8**.

30. Brunk, E., et al., *Recon3D enables a three-dimensional view of gene variation in human metabolism.* Nature Biotechnology, 2018. **36**(3): p. 272-+.

31. Robinson, J.L., et al., *An atlas of human metabolism.* Science Signaling, 2020. **13**(624).

32. Lewis, N.E., H. Nagarajan, and B.O. Palsson, *Constraining the metabolic genotype-phenotype relationship using a phylogeny of in silico methods.* Nature Reviews Microbiology, 2012. **10**(4): p. 291-305.

33. Orth, J.D., I. Thiele, and B.O. Palsson, *What is flux balance analysis?* Nature Biotechnology, 2010. **28**(3): p. 245-248.

34. Soh, K.C.a.H., V., *Constraining the flux space using thermodynamics and integration of metabolomics data.* Methods Mol Biol (Clifton, N.J.), 2014(1191): p. pp. 49-63.

35. Henry, C.S., L.J. Broadbelt, and V. Hatzimanikatis, *Thermodynamics-based metabolic flux analysis.* Biophysical Journal, 2007. **92**(5): p. 1792-1805.

36. Agren, R., et al., *Identification of anticancer drugs for hepatocellular carcinoma through personalized genome-scale metabolic modeling.* Mol Syst Biol, 2014. **10**: p. 721.

37. Wang, Y., J.A. Eddy, and N.D. Price, *Reconstruction of genome-scale metabolic models for 126 human tissues using mCADRE.* BMC Syst Biol, 2012. **6**: p. 153.

38. Ataman, M., et al., *redGEM: Systematic reduction and analysis of genome-scale metabolic reconstructions for development of consistent core metabolic models.* Plos Computational Biology, 2017. **13**(7).

39. Ataman, M. and V. Hatzimanikatis, *lumpGEM: Systematic generation of subnetworks and elementally balanced lumped reactions for the biosynthesis of target metabolites.* Plos Computational Biology, 2017. **13**(7).

40. Pacheco, M.P. and T. Sauter, *The FASTCORE Family: For the Fast Reconstruction of Compact Context-Specific Metabolic Networks Models.* Methods Mol Biol, 2018. **1716**: p. 101-110.

41. Nielsen, J., *Systems Biology of Metabolism: A Driver for Developing Personalized and Precision Medicine.* Cell Metabolism, 2017. **25**(3): p. 572-579.

42. Gu, C., et al., *Current status and applications of genome-scale metabolic models.* Genome Biol, 2019. **20**(1): p. 121.

43. Mendoza, S.N., et al., *A systematic assessment of current genome-scale metabolic reconstruction tools.* Genome Biol, 2019. **20**(1): p. 158.

44. Hughey, J.J., T.K. Lee, and M.W. Covert, *Computational modeling of mammalian signaling networks.* Wiley Interdisciplinary Reviews-Systems Biology and Medicine, 2010. **2**(2): p. 194-209.

45. Le Novere, N., *Quantitative and logic modelling of molecular and gene networks.* Nature Reviews Genetics, 2015. **16**(3): p. 146-158.

46. Hyduke, D.R. and B.O. Palsson, *Towards genome-scale signalling-network reconstructions.* Nature Reviews Genetics, 2010. **11**(4): p. 297-307.

47. Calzone, L., E. Barillot, and A. Zinovyev, *Predicting genetic interactions from Boolean models of biological networks.* Integrative Biology, 2015. **7**(8): p. 921-929.

48. Morris, M.K., et al., *Logic-Based Models for the Analysis of Cell Signaling Networks.* Biochemistry, 2010. **49**(15): p. 3216-3224.

49. Eungdamrong, N.J. and R. Iyengar, *Modeling cell signaling networks.* Biology of the Cell, 2004. **96**(5): p. 355-362.

50. Henriques, D., et al., *Data-driven reverse engineering of signaling pathways using ensembles of dynamic models.* Plos Computational Biology, 2017. **13**(2).

51. Covert, M.W., C.H. Schilling, and B. Palsson, *Regulation of gene expression in flux balance models of metabolism.* Journal of Theoretical Biology, 2001. **213**(1): p. 73-88.

52. Chandrasekaran, S. and N.D. Price, *Probabilistic integrative modeling of genome-scale metabolic and regulatory networks in Escherichia coli and Mycobacterium tuberculosis.* Proceedings of the National Academy of Sciences of the United States of America, 2010. **107**(41): p. 17845-17850.

53. Wang, Z., et al., *Combining inferred regulatory and reconstructed metabolic networks enhances phenotype prediction in yeast.* PLoS Comput Biol, 2017. **13**(5): p. e1005489.

54. Marmiesse, L., R. Peyraud, and L. Cottret, *FlexFlux: combining metabolic flux and regulatory network analyses.* Bmc Systems Biology, 2015. **9**.

55. Motamedian, E., et al., *TRFBA: an algorithm to integrate genome-scale metabolic and transcriptional regulatory networks with incorporation of expression data.* Bioinformatics, 2017. **33**(7): p. 1057-1063.

56. O'Sullivan, D., et al., *Metabolic interventions in the immune response to cancer.* Nat Rev Immunol, 2019. **19**(5): p. 324-335.

Chapter 2 Analysis of human metabolism and growth media using reduced thermodynamically curated genome-scale models.

In this chapter, we present a thermodynamic curation of the human genome-scale metabolic models and a novel workflow (redHUMAN) to characterize the extracellular medium in the models and to reconstruct reduced models that focus on a set of pathways that are of interest for the physiology under study. The method of this chapter has been developed in collaboration with Dr. M. Ataman. The content of this chapter has been published in *Nature Communications.*

2.1 Introduction

An altered metabolism is a hallmark of several human diseases, such as cancer, diabetes, obesity, Alzheimer's, and cardiovascular disorders [1, 2]. Understanding the metabolic mechanisms that underlie this reprogramming guides the discovery of new drug targets and the design of new therapies. To this effect, tremendous efforts are now being made to use the large amounts of now-available multi-omics experimental data to gain insight into the metabolic alterations occurring in different phenotypes.

Unfortunately, current mathematical models can be too complex for this analysis, rendering them too cumbersome to employ for many systems biology studies.

In the field of systems biology, genome-scale metabolic models (GEMs) integrate available omics data with genome sequences to provide an improved mechanistic understanding of the intracellular metabolism of an organism. GEMs have been reconstructed for a large diversity of organisms spanning from bacteria to mammals [3-5] and are valuable tools for studying metabolism [6, 7]. The mathematical representation of GEMs through the stoichiometric matrix [7] is amenable to methods such as flux balance analysis (FBA) [8] and thermodynamic-based flux balance analysis (TFA) [9-13], which ensure that the modeled metabolic reactions retain feasible concentrations and their directionalities obey the rules of thermodynamics, to predict reaction rates and metabolite concentrations when optimizing for a cellular function, such as growth, energy maintenance, or a specific metabolic task. Additionally, GEMs can be used for gene essentiality [14], drug off-target analysis [15], metabolic engineering [16-18], and the derivation of kinetic models [19-22].

The first human GEM was reconstructed in 2007 [23, 24]. Since then, the scientific community has been working to develop high-quality human GEMs, including HMR 2.0 [25], Recon 2 [26], Recon 2.2 [27], and Recon 3D [28]. The human GEMs used for the analysis in this chapter are Recon 2 and Recon 3D. Recon 2 is composed of 7440 reactions with 4821 of them associated to 2140 genes, and 2499 unique metabolites across seven compartments: cytosol, mitochondria, peroxisome, Golgi apparatus, endoplasmic reticulum, nucleus, and lysosome. Recon 3D is the latest consensus human GEM. It is an improved more comprehensive version of the previous GEMs consisting of 10600 reactions, with 5938 of them associated with 2248 genes, and 2797 unique metabolites compartmentalized as Recon 2 with an additional compartment for the mitochondria intermembrane space.

Human GEMs reconstruct the metabolic reactions occurring in several human cell types. However, a given cell type only leverages a portion of these reactions. This motivates the development of methods to generate context-specific metabolic models that can be used to study the differences in metabolism for different cell types [29], for healthy and diseased cells [30, 31], and for cells growing under diverse extracellular conditions. Some examples of such methods are: (1) GIMME [32], mCADRE [33] and tINIT [34] to

reconstruct tissue-specific models based on *omics* data and a set of tasks or a specific objective function; (2) redGEM-lumpGEM [35, 36] to reconstruct models around a specific set of subsystems of interest for the study; and (3) iMM [37, 38] to characterize the extracellular medium and the metabolites that are essential for growth under each condition. Context-specific metabolic models have been extensively used to understand the differences in metabolism between cancer cells and their healthy counterparts [39-45].

In this chapter, we present redHUMAN, a workflow to reconstruct novel thermodynamic-curated reductions of the human GEMs Recon 2 and Recon 3D. We integrate the thermodynamic properties of the metabolites and reactions into the GEMs and use redGEM-lumpGEM to reconstruct reduced models around specific subsystems. Furthermore, we introduce redGEMX, a method to identify the pathways required to connect the extracellular compounds to a core network. redGEMX guarantees that the reduced models have all the feasible pathways that consume and produce the components of the extracellular environment of the cell. Finally, we use metabolic data for leukemia as an example of how to integrate experimental data to derive disease- and tissue-specific metabolic models.

2.2 Results

2.2.1 Overall workflow

In order to generate reduced models from human GEMs, we developed redHUMAN, a six-step workflow that can be applied to any GEM or desired model system. The overall workflow is briefly described here and shown in Figure 2.1, and the details of each step in its application to the human GEMs Recon 2 and Recon 3D to generate thermodynamic-curated reductions are provided in the subsequent sections. For the workflow, the thermodynamic information for compounds and reactions, which is assembled from earlier studies or estimated using established group contribution methods, is first integrated into the GEM. Second, the subsystems, or families of pathways with a specific functional role for a biological process, are selected based on the objectives of the specific study. These pathways are explicitly represented and constitute the core of the reduced model. For example, when studying cancer

metabolism, this can include reported subsystems that are deregulated in cancer cells in addition to the standard central carbon pathways. Third, these subsystems are expanded using reactions from the GEM to create a connected core network. In this step, we include every reaction that connects core metabolites and that is not a member of the formal definition of the selected subsystems in the core model. In steps four and five, we include the shortest pathways to connect the extracellular metabolites from the defined medium as well as the shortest pathways to generate the biomass components from the core network. These steps guarantee that the model has all pathways that are essential for survival and growth of the cells based on the availability of nutrients. In the sixth step, experimental data for a specific physiological state is integrated in the model, and the final model is verified through checks that ensure the consistency of the reduced model with the original GEM.

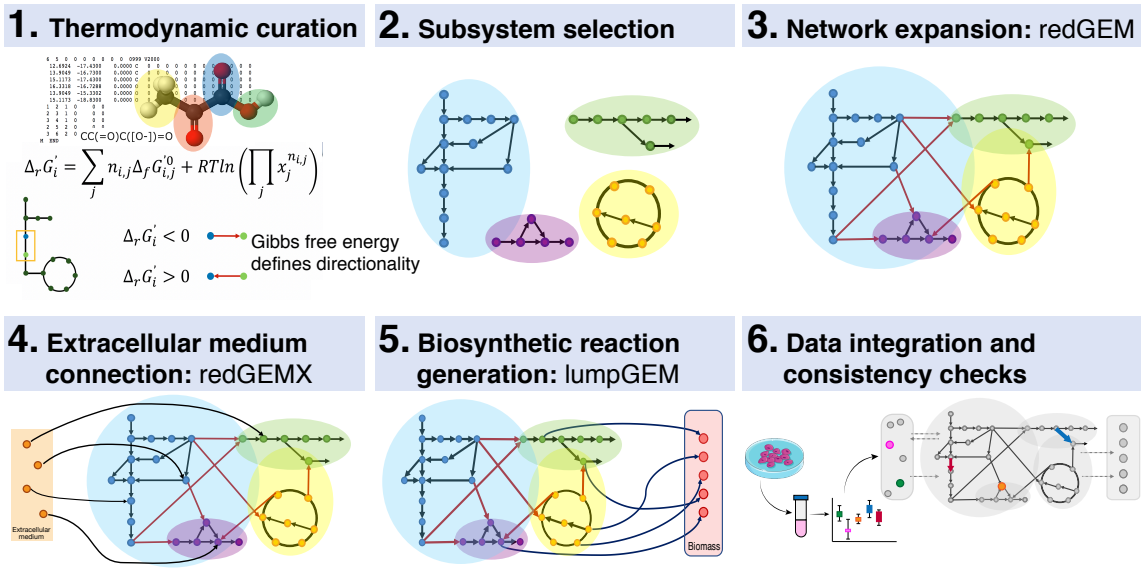


Figure 2.1. **redHUMAN: workflow to systematically reconstruct thermodynamic-curated reduced models.** (1) Thermodynamic curation: the Gibbs free energy of compounds and reactions are estimated and used to define the reaction directionality. (2) Subsystem selection: the subsystems relevant for the study are selected. (3) Network expansion: the initial subsystems are connected using reactions from the GEM to generate a core network. (4) Extracellular medium connection: the pathways that connect the extracellular medium components to the core network are identified. (5) Biosynthetic reaction generation: the pathways required to produce the biomass building blocks are classified. (6) Data integration and consistency checks: experimental values are integrated and the model is verified through consistency checks.

2.2.2 Thermodynamic curation of the human GEMs (Step 1)

We first determine the directionality of the chemical reactions of the network, which is directly associated with their corresponding Gibbs free energy. The Gibbs free energy of a reaction can be estimated from the thermodynamic properties of its reactants and products. Therefore, we curated the GEMs Recon 2 and Recon 3D (Materials and Methods) and integrated the thermodynamic properties for 52.4% of the 2499 unique metabolites from Recon 2 and 67.5% of the 2797 unique metabolites from Recon 3D (Figure 2.2 A). Three main reasons prevented the estimation of the thermodynamic properties of the metabolites: (1) an unknown molecular structure (SMILE), (2) an incomplete elemental description (for example, an R in the structure), and (3) groups in the structure for which an estimated free energy does not exist (for example, $>N^-$ group). We observed that as the number of metabolites increases from Recon 2 to Recon 3D, the percentage of thermodynamic coverage increases as well. This is due to the improved annotation of the metabolite structures in Recon 3D. Using the thermodynamic properties of the compounds as constraints (Materials and Methods), we estimated the Gibbs free energy for 51.3% of the 7440 reactions present in Recon 2 and 61.6% of the 10600 reactions in Recon 3D. These constraints ensured that the reactions in the computed flux distributions operated in thermodynamically feasible directions.

2.2.3 Subsystem selection to build the core (Step 2)

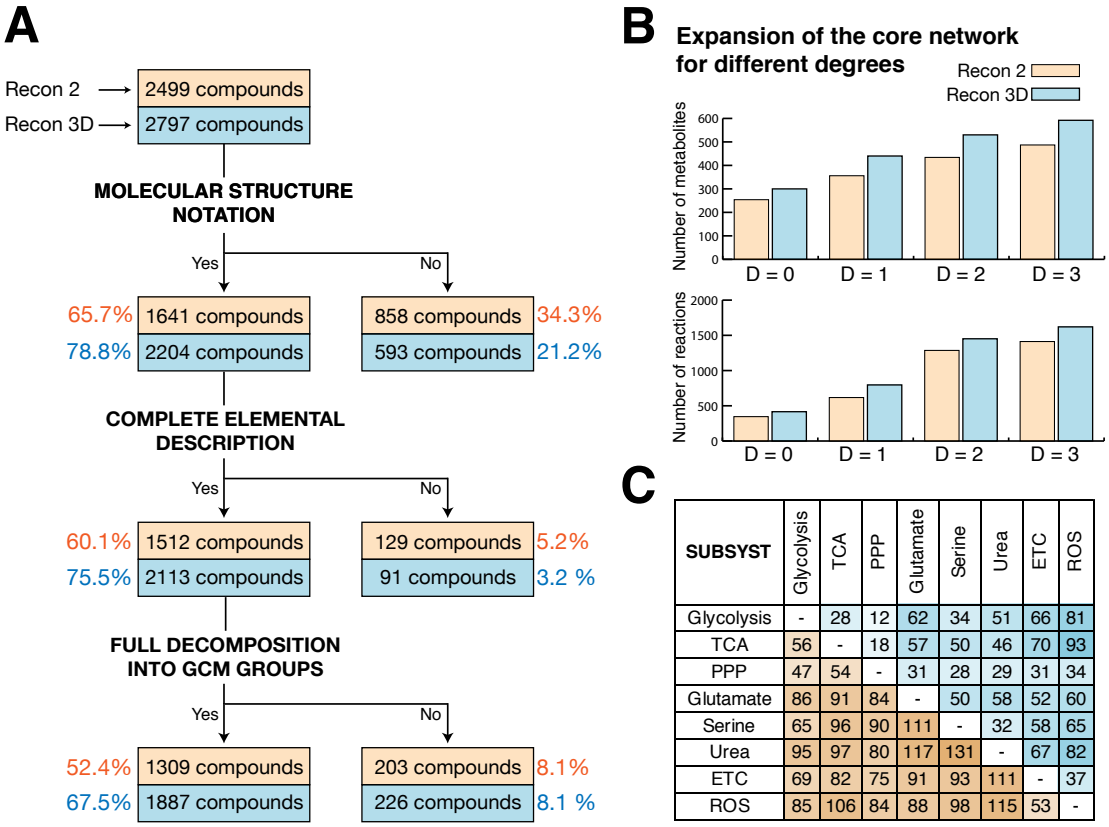
A proper metabolic model contains the pathways that are essential for the survival of the cell as well as the pathways that are informative of a specific metabolic behavior. In this work, we were interested in the metabolism of cancer cells. Thus, we selected as core subsystems: (a) the central carbon pathways that provide energy, the redox potential, and biomass precursors, and (b) the subsystems that have been reported to be altered in cancer cells [46-49]. Consequently, the core subsystems for our models were glycolysis, the pentose phosphate pathway, the citric acid cycle, oxidative phosphorylation, glutamate metabolism, serine metabolism, the urea cycle, and reactive oxygen species (ROS) detoxification. We have estimated the thermodynamic properties for the metabolites and the reactions in these initial subsystems. In the case of Recon 2 we provide an estimate for the Gibbs free energy of formation for 236 metabolites (94.4% of the total in the initial subsystems) and the Gibbs free energy of reaction for 143 reactions (83.1% of the reactions in the initial subsystems). In the case of Recon 3D, we

provide estimated values of the thermodynamic properties for 288 metabolites (97.6%) and for 183 reactions (91.0%).

2.2.4 Network expansion (Step 3)

Subsequently, to reconstruct the core network we pairwise connected the chosen subsystems using redGEM (Materials and Methods). The algorithm first performed an intra-expansion of the initial subsystems. In this process, each initial subsystem was expanded to include additional reactions from the GEM whose reactants and products belong to that subsystem. These reactions can be assigned to different subsystems in the GEM which are not any of the initial subsystems and the core network would miss these additional reactions if we had considered the formal definition of the initial subsystems. The initial core subsystems of Recon 2 contained a total of 180 reactions. After the intra-expansion, 135 reactions from 21 subsystems were added. Examples of these added reactions included three from *pyruvate metabolism* that interconvert acetyl-coa, acetate, malate, and pyruvate, which are all metabolites that participate in the *citric acid cycle* subsystem. For Recon 3D, 171 reactions from 24 subsystems were added to the 211 reactions from the initial core subsystems.

Next, the algorithm performed a directed graph search to find the reactions from the GEM that connected the subsystems for different degrees D (Figure 2.2 B and Table S2.1), wherein D represents the distance (in number of reactions) between pairs of metabolites from the subsystems. Our final models included the connections for degree D = 1, that is, all the reactions that in one step connect two metabolites (excluding cofactors) belonging to any of the initial subsystems. A degree D = 1 was enough to pairwise connect all the initial subsystems (Figure 2.2 C). This resulted in a Recon 2 core network of 356 metabolites and 617 reactions and a Recon 3D core network of 440 metabolites and 796 reactions.



SUBSYST	Glycolysis	TCA	PPP	Glutamate	Serine	Urea	ETC	ROS
Glycolysis	-	28	12	62	34	51	66	81
TCA	56	-	18	57	50	46	70	93
PPP	47	54	-	31	28	29	31	34
Glutamate	86	91	84	-	50	58	52	60
Serine	65	96	90	111	-	32	58	65
Urea	95	97	80	117	131	-	67	82
ETC	69	82	75	91	93	111	-	37
ROS	85	106	84	88	98	115	53	-

Figure 2.2. **Thermodynamic curation of human GEMs**. (A) Thermodynamics for the unique compounds in Recon 2 (orange) and Recon 3D (blue). The percentage is relative to the total number of unique compounds. (B) Size of the core network when the expansion is performed for different degrees. (C) Number of reactions that pairwise connect the subsystems for Recon 2 (values below the diagonal) and Recon 3D (values above the diagonal) for degree D = 1.

2.2.5 Extracellular medium connection (Step 4)

Cells adapt their metabolism to the available nutrients in their extracellular environment. Consequently, a correct definition of the medium in the metabolic model is fundamental for an adequate representation of the intracellular metabolism. Given the complexity of the extracellular medium, it is particularly important to identify and classify the essentiality of the medium components and the pathways used for their metabolism. To this end, we curated the representation of the interactions of the cell with its environment into the human GEMs. First, we did not allow the exchange of intracellular metabolites lacking associated transport reactions or transport molecules containing P, CoA, or ACP. Secondly, we allowed the synthesis of generic fatty acids from palmitate, with reactions from Recon 2 and Recon 3D (Note S2.1). We next characterized the *in silico* minimal medium composition required for growth in the human GEMs by applying iMM (Materials

and Methods), which identifies the minimal set of metabolites that need to be uptaken to simulate growth. The results showed that Recon 2 required a medium with glucose, the nine essential amino acids, and some inorganics (PO_4, NH_4, SO_4, O_2), and Recon 3D simulated growth in a medium with glucose, the nine essential amino acids, the same inorganics as Recon 2, and one of the two essential fatty acids (alpha-linolenic acid and linoleic acid). The presence of the two essential fatty acids in the iMM of Recon 3D is a consequence of the improvement of the lipid metabolism [28], where the essential fatty acids participate in the synthesis of phospholipids. This demonstrates how the algorithms and workflow can be used to compare and validate updated model reconstructions for the same organisms or between different organisms.

Seeking to identify the pathways that human cells use to uptake and secrete extracellular metabolites, we next developed the method redGEMX (Materials and Methods). This algorithm finds the pathways from the GEM that are needed to connect the extracellular metabolites to the core network defined by redGEM. In this work, we considered a complex medium composition of 34 metabolites (Figure 2.3 A), and redGEMX found the corresponding reactions from the GEM that connected 26 of these extracellular metabolites (we excluded the inorganics and the fatty acids) to the core network.

A

Extracellular medium

Amino Acids	L-glutamate	L-arginine
	L-glutamine	L-aspartate
	L-proline	L-asparagine
	L-serine	L-lysine
	L-glycine	L-methionine
	L-alanine	L-threonine
	L-valine	L-tryptophan
	L-isoleucine	L-tyrosine
	L-histidine	L-phenylalanine
	L-leucine	
Central carbon	D-glucose	L-malate
	L-lactate	Citrate
Inorganics & gases	Oxygen	Carbon dioxide
	Water	Ammonium
	Phosphate	Sulfate
Urea cycle	Urea	Ornithine
Vitamin	Choline	
Fatty Acids	Linoleic acid	
	α-linolenic acid	

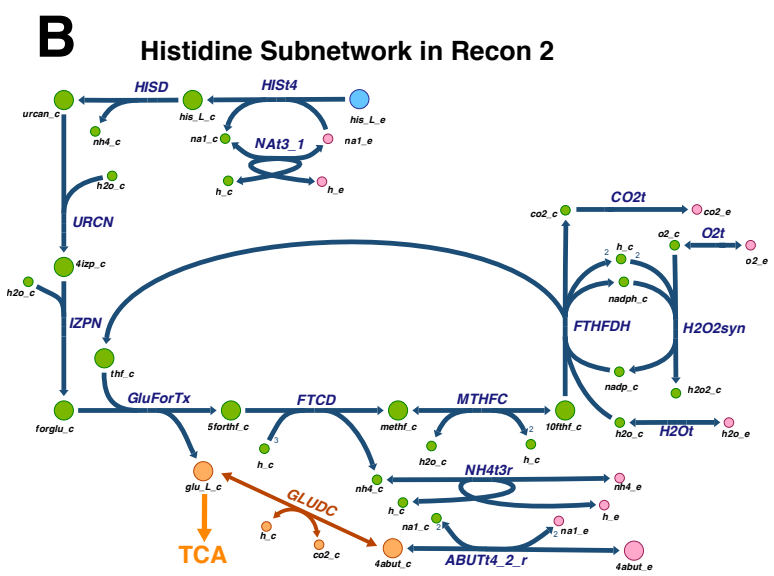


Figure 2.3. **Extracellular medium utilization** (A) Extracellular medium composition defined in the models. (B) Graph of the subnetwork from Recon 2 for the uptake of L-histidine and the medium components required for its metabolism. Green represents the metabolites from the subnetwork, and orange represents the metabolites of the core network where the subnetwork is connected. In blue, the medium metabolite under study (L-histidine) and in pink, the extracellular metabolites co-utilized to metabolize L-histidine. The pathway starts with the transport of L-

histidine from the extracellular space to the cytosol, where it is sequentially transformed into urocanate (urcan_c), 4-imidazolone-5-propanoate (4izp_c), N-formimidoyl-L-glutamate (forglu_c), L-glutamate (glu_L), 5-formiminotetrahydrofolate (5forthf_c), 5-10-methenyltetrahydrofolate (methf_c), and 10-formyltetrahydrofolate (10thf_c). 4-Aminobutanoate (4abut_c) is converted to L-glutamate through a reaction from the subsystem glutamate metabolism, and finally, L-glutamate is connected to the TCA cycle.

An example of one of these connected metabolites is the essential amino acid L-histidine which affects many aspects of human physiology, including cognition functions and allergic reactions. The classical pathway to metabolize L-histidine consists of four steps that sequentially convert it into urocanate, 4-imidazole-5-propanoate, N-formimidoyl-L-glutamate, and ultimately, L-glutamate [50]. Interestingly, the resulting redGEMX subnetwork for L-histidine uses this classical pathway to connect it to the Recon 2 core metabolites L-glutamate and 4-aminobutanoate, both from the subsystem *glutamate metabolism*. The subnetwork is composed of 22 reactions, and it contains not only the classical pathway but also all the additional reactions required to balance the cofactors and by-products (Figure 2.3 B). These additional reactions are essential for an active main pathway, as they include the utilization of NH_4, the sources of water and tetrahydrofolate, and the conversion of the by-product 5-formiminotetrahydrofolate to 10-formyltetrahydrofolate, which regenerates tetrahydrofolate. Cellular metabolism has evolved to give flexibility to the cells to survive and function under different conditions. This flexibility is captured in the metabolic networks with the existence of alternative pathways. For this reason, using redGEMX we found three alternative pathways of minimum size (22 reactions) to connect L-histidine to the core network of Recon 2. The alternatives emerge from the existence of different transport reactions for the extracellular metabolites. In the case of Recon 3D, L-histidine is connected to the core network using 20 reactions, and there exist two pathways of minimum size. The subnetworks connect L-histidine to the Recon 3D core metabolites glutamate, 5-10-methylenetetrahydrofolate, 2-oxoglutarate and pyruvate using the classical pathway to metabolize L-histidine. The different topology of Recon 2 and Recon 3D networks manifests in differences in the pathways used to metabolize and synthesize the compounds, thus, it is important to characterize which are the pathways used in the models. Following this approach, we added the reactions that compose all the alternative subnetworks of minimum size to the core networks to connect the 26 extracellular metabolites (Table S2.2).

The subnetworks generated with redGEMX provide a new perspective on the current understanding of metabolic pathways, as they not only provide the main pathway but they also include other reactions necessary to provide and consume all the by-products and cofactors. Moreover, the alternatives can be used to hypothesize which pathways cells use when growing under different conditions, such as when different nutrients are present in the environment or under different intracellular regulations when different enzymes are operational. If metabolomics data are available, the subnetworks generated with redGEMX can be classified based on pathway favorability as it has been recently done in [9, 51, 52].

2.2.6 Biosynthetic reactions generation (Step 5)

Cellular metabolic functions, such as growth, structure maintenance, and reproduction, require the synthesis of several metabolites. In metabolic models, this is represented using the biomass reaction [53], whose reactants, named biomass building blocks or BBBs, are the metabolites that the cell needs to survive and perform its functions. Therefore, the last step necessary for reconstructing the reduced models is the integration of the pathways necessary to synthesize the 37 BBBs that compose the defined biomass in Recon 2 and Recon 3D. Among them, 19 are uptaken directly from the extracellular medium or produced within the core network. To find the minimum number of reactions in the GEM that we need to add to the core network for the synthesis of the remaining 18 BBBs, we used lumpGEM (Materials and Methods). Similarly to redGEMX, lumpGEM generates subnetworks that account for the synthesis, degradation, and balancing of all the by-products and cofactors required by the main pathway. The alternative subnetworks generated with lumpGEM can assess the flexibility of the cells to use alternative pathways to produce the BBBs, which can lead to survival in different conditions and drug resistance. Using lumpGEM, we calculated all the alternative subnetworks (set of reactions) of minimum size to capture the flexibility of the network for the biosynthesis of the BBBs (Figure 2.4 A, Table S2.3). The reactions that compose each of these subnetworks were summed up together to form an overall reaction that represented the subnetwork. These lumped reactions were then added to the core network.

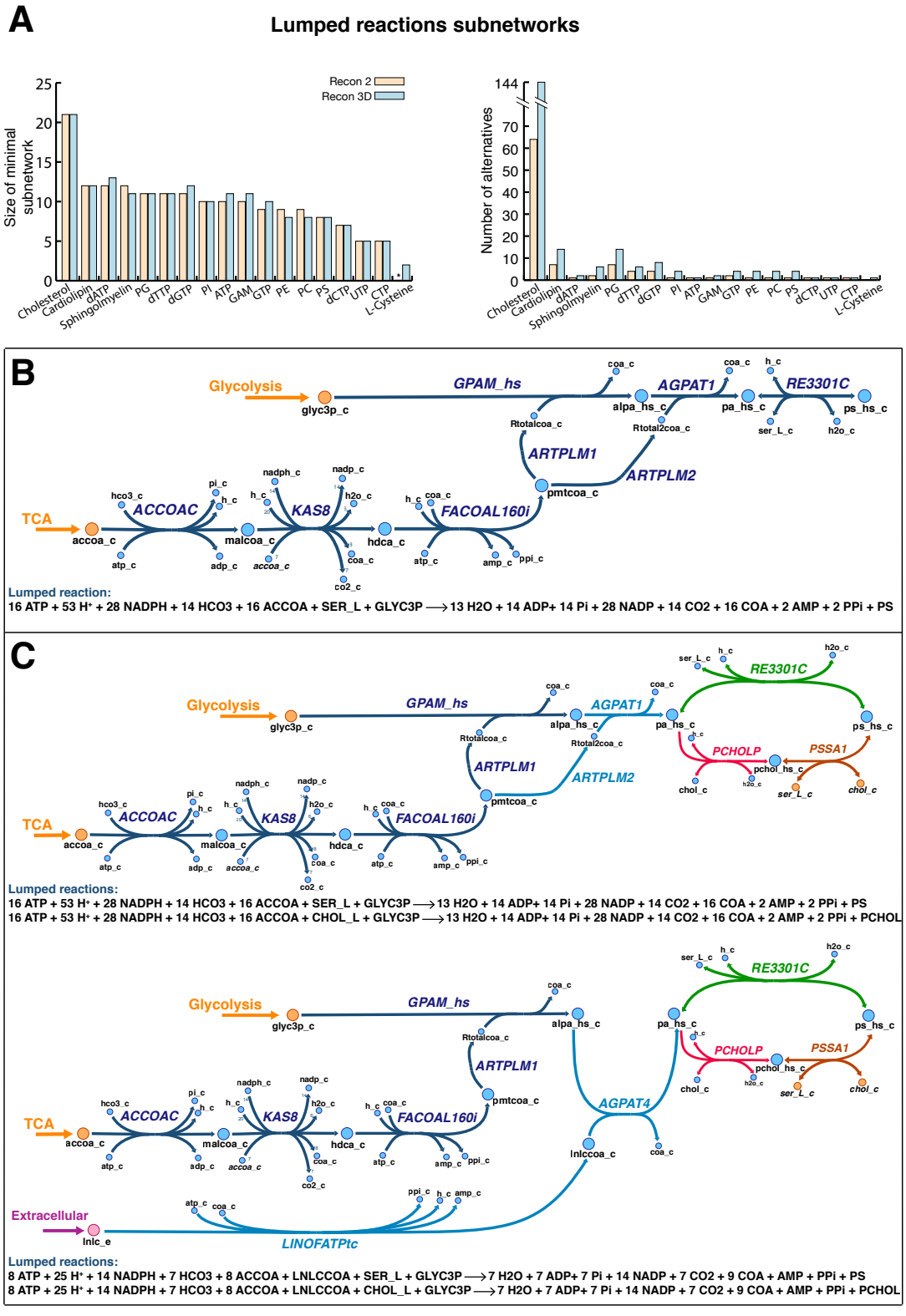


Figure 2.4. **Biosynthesis of biomass building blocks.** (A) Size of lumped reactions for Recon 2 and Recon 3D, and the corresponding number of alternatives to synthesize the BBBs that cannot be produced by the core nor uptaken from the extracellular medium. (B-C) Subnetwork for the synthesis of phosphatidylserine. Orange represents the metabolites from the core network. Blue represents the metabolites from the subnetwork for phosphatidylserine synthesis. Pink

represents the extracellular metabolites. Phosphatidylserine synthesis starts from the core metabolites glycerol 3-phosphate, from glycolysis, and acetyl CoA, from TCA. In a first reaction, acetyl CoA is transformed into malonyl CoA. The next reaction (KAS8) represents the synthesis of palmitate in the elongation cycle [54]. A CoA molecule is attached to palmitate to form palmitoyl CoA, from which the two generic fatty acids are derived. These two generic fatty acids are attached to glycerol 3-phosphate to form lysophosphatidic acid and phosphatidic acid. Finally, serine is attached to phosphatidic acid to form phosphatidylserine (B) Subnetwork from Recon 2 and corresponding lumped reaction. (C) The four alternative subnetworks of minimum size from Recon 3D. Phosphatidic acid can be produced with two generic fatty acids or with one generic fatty acid and the essential fatty acid linoleic acid (light blue reactions). Phosphatidylserine can be directly produced from phosphatidic acid by attaching serine (green reaction) or through the formation of phosphatidylcholine (red reaction) and then changing choline for serine (orange reaction).

The subnetworks generated with lumpGEM have the same size and number of alternatives in both Recon models for most of the BBBs, indicating that both models have the same level of flexibility for synthesizing the BBBs, with the exception of L-cysteine, dTTP and the purine nucleotides (ATP, GTP and their deoxy equivalents), cholesterol, and the phospholipids and sphingolipids. The core network of Recon 2 contains a reaction that produces L-cysteine, however, the core network of Recon 3D requires two reactions to produce it. The subnetworks that produce dTTP have the same size in both models, but a different number of alternatives. The subnetworks to produce the purine nucleotides have one more reaction and more alternatives in Recon 3D. Cholesterol is another BBB whose subnetworks agree in size for both models, but Recon 3D has more alternatives than Recon 2. The explosion of alternatives in Recon 3D is due to the parallel description of the synthesis of cholesterol in three compartments, namely cytosol, peroxisome, and endoplasmic reticulum. The differences in the lumped reactions for the phospholipids and sphingolipids between both models are due to the introduction of the essential fatty acid in their synthesis in Recon 3D.

As an example of the subnetworks that produce the BBBs, we show the synthesis of the phospholipid phosphatidylserine (Figure 2.4 B and C). The standard KEGG pathway [55] for the synthesis of phosphatidylserine comprises four steps, wherein glycerol 3-phosphate is converted to lysophosphatidic acid, phosphatidic acid, CDP-diacylglycerol, and phosphatidylserine. In Recon 2, the subnetwork generated with lumpGEM for the synthesis of phosphatidylserine was composed of eight reactions. It included the KEGG pathway with the exception of the CDP-diacylglycerol intermediate, which was not connected to phosphatidylserine in the GEMs. Instead, phosphatidylserine was

produced directly from phosphatidic acid by attaching serine. Additionally, the subnetwork contained the reactions required to generate from acetyl-CoA the fatty acids that would attach to glycerol 3-phosphate and to lysophosphatidic acid, which are important to consider for the final synthesis of phosphatidylserine. All the reactions involved in the synthesis of phosphatidylserine were lumped together in one reaction.

For Recon 3D, the phosphatidylserine synthesis subnetwork was generated with the same eight reactions, but in this case, four alternative subnetworks existed (Figure 2.4 C and Table S2.4), indicating that Recon 3D has a higher flexibility in producing this BBB. The alternatives emerged from the presence of two reactions in Recon 3D that could be substituted by two other reactions in the subnetwork. One of these reactions arose from the participation of the essential fatty acid linoleate in phospholipid generation, resulting in an alternative form of synthesizing one of the tails of phosphatidic acid. Specifically, the reaction ARTPLM2, which converts palmitoyl CoA into a generic fatty acid, is not required, and instead, the essential fatty acid linoleate is transported from the extracellular medium, transformed into linoleyl-coA and attached to the lysophosphatidic acid to form phosphatidic acid. Because the core network of Recon 3D included a reaction that transforms phosphatidylcholine in phosphatidylserine, the other substitution occurred in the last step, where serine was replaced by choline and phosphatidylcholine was synthesized. The lumped reactions can be classified based on the thermodynamic favorability of their subnetworks, if metabolomics data are available, as in [9, 51, 52].

The analysis performed with lumpGEM allows to characterize and classify the metabolic pathways and their alternatives, leading to an in-depth understanding of the flexibility of metabolism. In the context of GEMs, such detailed analysis of the subnetworks is often a difficult task due to their large size and interconnectivity.

By applying the redHUMAN workflow, we reconstructed four reduced metabolic models for human metabolism (Table 2.1). Two of them have Recon 2 as the parent GEM, and the other two are generated from the Recon 3D GEM. For both GEMs, we generated one model with the minimum set of pathways required to simulate growth, that is, one lumped reaction per BBB with subnetworks of minimum size, and another model with higher flexibility containing all the alternative pathways of minimum size required to simulate growth. The reduced models have a thermodynamic coverage of more than 92% of the compounds and more than 61% of the reactions.

Table 2.1**. Statistics on the generated reduced metabolic models**. The models were generated from the human GEMs Recon 2 and Recon 3D. For each GEM, two reductions were performed considering either one lumped reaction per BBB (OPBBB) or all the alternatives lumped reaction with subnetworks of minimum size (Smin).

		Recon 2			**Recon 3D**		
Model		GEM	Reduced Recon 2 OPBBB	Reduced Recon 2 Smin	GEM	Reduced Recon 3D OPBBB	Reduced Recon 3D Smin
Number of metabolites		5063	469	469	5835	591	599
Num. of reactions	Enzymatic	4220	342	342	4609	402	405
	Boundary	701	71	71	1863	130	130
	Transports	2519	946	946	4187	1085	1092
	Lumped	-	15	37	-	15	105
	TOTAL	**7440**	**1374**	**1396**	**10602**	**1632**	**1732**
Number of genes		2140	699	699	2248	747	748
% of metabolites with est. Gibbs energies		61.1	92.8	92.8	71.1	93.7	93.8
% of reactions with est. Gibbs energies		51.3	62.3	61.7	61.6	63.5	62.0

2.2.7 Data integration and metabolic tasks (Step 6)

Once the reduced models were generated, we investigated the metabolic tasks captured by the reduced models and we identified how the models should be curated to recover the tasks that they could not perform. First, we sequentially tested in the generated reduced models the thermodynamically feasibility of 57 metabolic tasks defined by Agren et al [34]. The four models captured 45 of the 57 tasks, including rephosphorylation of nucleoside triphosphates, uptake of essential amino acids, de novo synthesis of nucleotides, key intermediates and cholesterol, oxidative phosphorylation, oxidative decarboxylation, and growth (Figure 2.5 A).

The tasks not captured by the models encompassed the synthesis of protein from amino acids, beta oxidation of fatty acids, inositol uptake, and vitamin and co-factor metabolism. We classified the causes behind their limitation into two categories: (1) the model

reconstruction, specifically the definition of the biomass, or (2) the reduction properties, that is, the subsystems included in the reduction and the representation of parts of the network as lumped reactions. To recover these tasks such that they are captured by the model, the following actions should be performed: the synthesis of proteins from amino acids and vitamin and co-factor metabolism can be recovered by modifying the biomass to account for their synthesis and utilization; the inclusion of lipid metabolism subsystems can recover the beta oxidation of fatty acids; and finally, the utilization of inositol can be recovered by adding the explicit reactions that compose the subnetworks, as it was found to be hidden in the lumped reactions of phosphatidyl-inositol. This demonstrates that redHUMAN allows to build reduced models consistent not only with the GEM but also with the metabolic tasks, and these models are suitable for targeted modifications and expansions.

We next demonstrated how generic reduced models were used to integrate data to study disease physiology. We first integrated experimental data from the NCI60 cell lines in the reduced models to define the physiology of leukemia cells. In particular, we considered the exometabolomics of the cell lines HL-60, K-562, MOLT-4, CCRF-CEM, RPMI-8226, and SR, which correspond to leukemia [40, 56]. Additionally, we limited the maximal growth to the doubling time reported for leukemia cells, which is 0.035 h^{-1}, and we constrained according to literature values the maximum uptake rate of oxygen to 2 mmol/gDW/h [40] and the ATP maintenance to 1.07 mmol/gDW/h [57] (Table S2.5 and Table S2.6). We tested that all the models achieved the maximum growth when maximizing for the biomass reaction using TFA.

Next, to analyze the impact that the deletion of each gene had on the network, we performed *in silico* gene knockout by artificially removing a gene and measuring how the network was affected. The genes whose knockout prevented the synthesis of biomass could then be investigated as potential targets for limiting cell proliferation. The consistency of the workflow used to generate the reduced models ensures that they capture the essentiality from the GEM, that is, the genes that are part of the reduced models and are essential in the GEM they are also essential in the reduced model (Figure 2.5 B, Table S2.7 and Table S2.8). Furthermore, the reduced models allow the assignment of functionality to the essential genes using the lumped reactions. For example, the gene GART is associated with the enzymes phosphoribosylglycinamide formyltransferase, phosphoribosylglycinamide synthetase, and

phosphoribosylaminoimidazole synthetase, which are all part of the subnetworks for the synthesis of the nucleotides ATP, GTP, dATP and dGTP. Silencing this gene prevents the synthesis of these BBBs, and consequently, the models cannot synthesize biomass. When information about synthetic lethality is available, we can test and validate the predicted essentiality against the experimental data.

Finally, because the model reduction affects the flexibility of the network with respect to the GEM, we performed thermodynamic flux variability analysis (TVA) on the common reactions between the GEM and the reduced model. The top 20 reactions whose rate ranges changed the most in absolute value included reactions from glycolysis, the pentose phosphate pathway, folate metabolism, and nucleotide interconversion among others (Figure 2.5 C). For reactions such as phosphoglycerate kinase (PGK), transaldolase (TALA) and methenyltetrahydrofolate cyclohydrolase (MTHFC), the ranges of reaction rates in the reduced model decreased with respect to the corresponding reaction rates in the GEM. Some reactions such as, nucleoside-diphosphate kinase (NDPK9), were bidirectional in the GEM and became unidirectional in the reduced models. On the other hand, there were also reactions such as fumarase (FUM), lactate dehydrogenase (LDHL), or ribose-5-phosphate isomerase (RPI) whose flux ranges fully agreed between the reduced model and the GEM. Interestingly, if we look at the percentage of rate flexibility change, the reactions from the initial subsystems did not experience a large relative change in their rates, with the exception of the reactions whose participants are precursors for the lumped reactions of the BBBs as their reaction rates are now constrained closer to the physiological state.

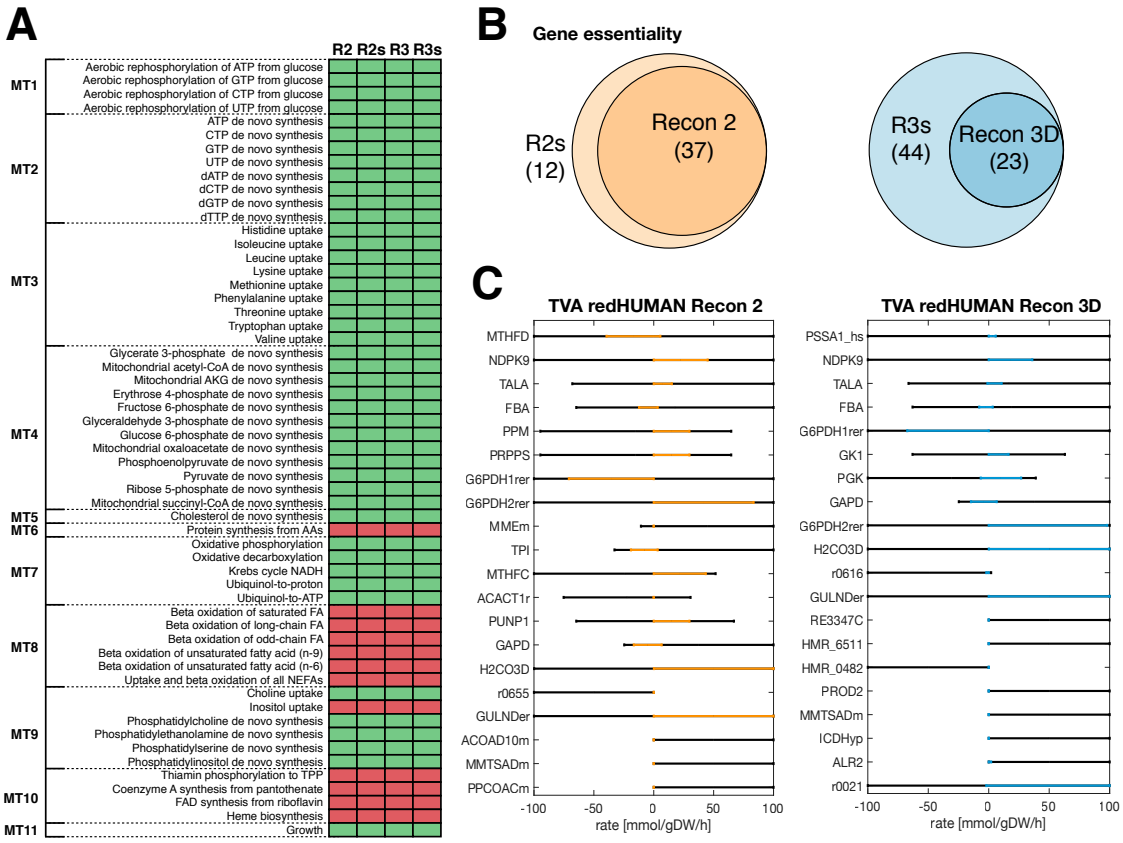


Figure 2.5. **Model validation through metabolic tasks and consistency checks**. (A) The 57 metabolic tasks tested in the generated reduced models. R2, R3: Recon 2, Recon 3D reduced model with one lumped reaction per BBB. R2s, R3s: Recon 2, Recon 3D reduced model with Smin. Classification of metabolic tasks in those captured by the models (green) and those not captured by the models (red). MT1: rephosphorylation of nucleoside triphosphates, MT2: de novo synthesis of nucleotides, MT3: uptake of essential amino acids, MT4: de novo synthesis of key intermediates, MT5: de novo synthesis of other compounds, MT6: protein turnover, MT7: electron transport chain and TCA, MT8: beta oxidation of fatty acids, MT9: de novo synthesis of phospholipids, MT10: vitamins and co-factors, MT11: growth. (B) Gene essentiality of the reduced models and their corresponding GEM. R2s has 829 genes associated to reactions, 37 of which are essential both in the reduced model and in Recon 2 and 12 are essential only in the reduced model. R3s has 828 genes associated to reactions, from which 23 are essential in both the reduced model and Recon 3D. The reduced model presents an additional 44 essential genes. (C) Thermo-flux variability analysis (TVA) for reactions in the reduced models. Orange represents fluxes in the reduced Recon 2 model and blue represents fluxes in the reduced Recon 3D models. The black lines correspond to the fluxes in the GEM.

A final calibration of the models is done using the transcriptomics data from the NCI data repository (https://www.ncbi.nlm.nih.gov/sites/GDSbrowser?acc=GDS4296) for the corresponding leukemia cell lines. We have identified that, in the four models presented in this study, over 99% of the enzymes with gene associations (more than 75% of the total enzymes) are expressed in the NCI60 leukemia cell lines (Table S2.9). This

suggests that the pathways selected for initializing and expanding the metabolic core network are highly relevant for the specific physiology, which are also consistent with the important pathways identified in the experimental and medical studies [46, 48, 58].

2.2.8 Physiology analysis

redHUMAN helps to navigate large human genome-scale metabolic models to explore and classify the metabolic pathways that cells use to function and survive under specific conditions. The thermodynamic curation performed in the genome-scale models guarantees that the reactions obey the laws of thermodynamics, discarding possible pathways that would not be compatible with the bioenergetics of the cell. As an example of how thermodynamics reduces the space of solutions to the thermodynamically feasible pathways, we analyzed the flux variability with and without thermodynamic constraints in the Recon 3D reduced model that has all the alternative lumped reactions of minimum size (Smin). The reactions L-Glutamate 5-Semialdehyde Dehydratase (G5SADs, from arginine metabolism) and L-Glutamate 5-Semialdehyde:NAD+ Oxidoreductase (r0074, from urea cycle) are bidirectional when flux variability is performed without thermodynamics and become unidirectional when their thermodynamic information is taken into account. Therefore, integrating thermodynamic information reduces the space of reaction directionality and the physiological solution space and eliminates thermodynamic infeasible reactions excluding some pathways.

The leukemia-specific models generated in this study are powerful tools to analyze how the metabolic pathways are altered with respect to other cancer cells or normal cells. In particular, we can analyze how leukemia cells utilize the nutrients available in the microenvironment to biosynthesize the precursors required for growth and cellular functionality. As an example, we identified the minimal number of reactions that are required for the synthesis of phosphatidyl-serine in the reduced Recon 3D model with all the alternative lumped reactions of minimum size. We found that at least 76 reactions should be active for the production of phosphatidyl-serine including the interactions with the extracellular medium, i.e., for some alternatives the uptake of glucose, histidine, linoleic acid, oxygen, and phosphate, and the secretion of succinate, ammonia, carbon dioxide and water. The main pathways active within the subnetwork of 76 reactions are glycolysis, the citric acid cycle, serine metabolism, and the electron transport chain. This type of analysis will enlighten our knowledge on how cells adapt their metabolism to the

microenvironment allowing researchers to hypothesize how and why the cancer cells change their expression profile to adapt and survive.

2.3 Discussion

For a better understanding of the altered metabolisms that accompany many human diseases, we have herein presented a workflow to generate reduced models for common human GEMs that can reduce the complexity of these systems to the relevant processes to be studied, making detailed in silico analyses of metabolic changes possible.

During the last years, there has been an increased generation of metabolomics data that better study what is happening in the physiology of cell metabolism compared to other omics data. This has created a need to expand the classical constraint-based modeling methods to include metabolomics information. Our thermodynamic formulation and application of TFA [12, 51, 59, 60] in redHUMAN allows to integrate endo- and exo-metabolomics in the models, constraining the concentration of the metabolites according to physiological data. The size of the model is directly related to the percentage of metabolites that need to be measured. Therefore, the continuous expansion in size of genome-scale models increases the demand of larger sets of metabolomics, and such data are not always available. In addition, there is a community effort to expand constraint-based models to include information on enzyme abundancy relating the metabolic fluxes with enzymatic data and allowing to integrate transcriptomics and proteomics data into the models. These data are currently limited but they can be continuously updated and integrated as they become available [61, 62].

Moreover, most of the existing methods to build context-specific models are data-driven, that is, the reduced models are extracted from a GEM by considering only the enzymes associated to highly expressed data, or literature-based pathways. Then, they include the additional reactions that are required to simulate growth and cellular functions [33, 34, 63]. The main difficulty with these methods is the large amount of data required to fully characterize the initial set of reactions, or core reactions. The lack of data could lead to unconnected parts and the impossibility to include reactions that could be important for the specific physiology, affecting the final model and the predictions.

redHUMAN reconstructs reduced models considering only the pathways of interest and their stoichiometric connectivity. The reduced models are built unbiased from the data, guaranteeing thermodynamic feasibility and consistency with the GEM and the metabolic tasks. The reduced models can then be used to construct context-specific models by integrating omics data, accommodating to also integrate partial data without sacrificing reactions from the network. Overall, the reduced size of the new models and their conceptual organization overcomes some of the main challenges in building genome-scale context-specific models as for example the barrier of data network coverage. The reduced models generated with redHUMAN are powerful representations of the specific parts of the network and have promising applications as they are suitable to use with existing methods including MBA [63], tINIT [34], mCADRE [33], uFBA [64], GECKO [65], ETFL [66], TEX-FBA [67] and IOMA [68].

Based on our results, we propose the following approach to using these models as tools to explain and compare phenotypes. First, generate a reduced model around a desired set of subsystems and for a defined extracellular medium, and check that the model captures the metabolic tasks. Subsequently, build physiology-specific models by integrating experimental data into the reduced models. Then, test the consistency of the reduced network with respect to its parent GEM. Finally, integrate different sets of omics data, including expression, to compare different physiologies, such as diseased vs healthy or within several types of cancers. This approach will help to better investigate the alterations in metabolism that occur as diseases develop and progress. Moreover, the same procedure can be used to analyze systematically and consistently metabolic models for the same organism and to compare metabolic models of different organisms, enhancing our understanding of their similarities and differences.

Throughout this chapter, we have considered a specific set of subsystems, a specific medium, and the biomass definition from the GEMs. In the future, the reduced models could be further expanded to include other pathways, a more complex medium, or more biomass components. To introduce new subsystems or pathways into the core network, redGEM should be run to find the pairwise connections between the added pathways and the rest of the core. For an expansion of the medium, redGEMX would find the connections necessary for using the new extracellular metabolites. In a similar manner, a further curation of the biomass reaction could increase the number of BBBs, requiring lumpGEM to be run to find the biosynthesis pathways for those compounds. If a higher

consistency was required between the GEM and the corresponding reduction, we could find the reactions missing in the reduced model to satisfy that condition.

Furthermore, in this study we have used metabolomics, proteomics and growth data from the NCI60 cell lines to define a generic physiology for leukemia cells. The core networks of the reduced models are structurally the same across growth conditions and depend only on the structure of the corresponding GEMs. Therefore, these generic models are robust to variations in growth or data for the same physiology and thus data for individual leukemia cell lines can be used without changing the workflow. However, if there are important differences in the data, for example across different physiological conditions, the authors suggest running the lumpGEM workflow with data integration and generate alternative subnetworks and lumped reactions, which in turn will capture the different flux profiles for each physiological state.

Overall, our analysis demonstrates how redHUMAN facilitates the characterization of differences in metabolic pathways across models and phenotypes.

2.4 Materials and Methods

2.4.1 Experimental data for leukemia cell lines

The experimental data used in this work are exo-metabolomics, exo-fluxomics and transcriptomics corresponding to the NCI60 leukemia-specific cell lines, HL-60, K-562, MOLT-4, CCRF-CEM, RPMI-8226, and SR.

We integrated in the models exo-metabolomics and exo-fluxomics for these cell lines measured and estimated by Jain et al. in previous work [56]. In particular, concentrations for 42 extracellular metabolite (Table S2.6) and reaction rates for 24 metabolites (Table S2.5) were used from the work by Jain et al. [56] as bounds for the corresponding variables in the models. Additionally, we constrained the uptake rate of oxygen to 2 mmol/gDW/h [40] and the ATP maintenance rate to 1.07 mmol/gDW/h [57]. Furthermore, the doubling time for leukemia cells was used to define the growth rate in the model (0.035 h^{-1}).

Transcriptomics data for metabolic genes from the NCI60 GDS4296 NCI data set were used to calibrate the models. The transcriptomics data corresponding to the leukemia cell lines was used to evaluate the expression of the genes present in the reduced model. The details regarding the cell lines, the raw data, and the processed data can be found in the NCI site https://www.ncbi.nlm.nih.gov/sites/GDSbrowser?acc=GDS4296.

2.4.2 Thermodynamic curation of the genome-scale models (GEMs)

The thermodynamic curation of the human GEMs Recon 2 and Recon 3D aims to include thermodynamic information, i.e., the Gibbs free energy of formation for the compounds and the corresponding error for the estimation, into the model. The workflow to obtain this information is as follows.

We first used MetaNetX (http://www.metanetx.org) [69] to annotate the compounds of the GEMs with identifiers from SEED [70], KEGG [55], CHEBI [71], and HMDB [72]. We then used Marvin (version 18.1, 2018, ChemAxon http://www.chemaxon.com) to transform the compound structures (canonical SMILES) into their major protonation states at pH 7 and to generate MDL Molfiles. We used the MDL Molfiles and the Group Contribution Method (GCM) to estimate the standard Gibbs free energy of the formation of the compounds as well as the error of the estimation [60].

Since the model for Recon 3D already incorporates the structure for 82% of the metabolites in the form of SMILES, we used those SMILES and followed the previous workflow from the point of obtaining the major forms at pH 7 using Marvin.

Furthermore, we have integrated in the models the thermodynamic properties for the compartments of human cells, including, pH, ionic strength, membrane potentials and generic compartment concentration ranges from 10pM to 0.1M (Table S2.10).

2.4.3 TFA: thermodynamics-based flux analysis

TFA estimates the feasible flux and concentration space according to the laws of thermodynamics [11-13]. TFA is formulated as a mixed-integer linear programming (MILP) problem that incorporates the thermodynamic constraints to the original flux balance analysis (FBA) problem. The Gibbs free energy of the elemental and charge balanced reactions is calculated as a function of the standard transformed Gibbs free energy of formation (depending on pH and ionic strength) and the concentrations of the products and reactants.

Considering a network with m metabolites and n reactions, the Gibbs free energy, $\Delta_r \mathrm{G}'_i$, for reaction i is:

$$\Delta_r \mathrm{G}'_i = \sum_{j=1}^{m} n_{i,j} \Delta_f \mathrm{G}'^{o}_j + RT \ln\left(\prod_{j=1}^{m} x_j^{n_{i,j}}\right),$$

where $i = 1, \dots, n,\ j = 1, \dots, m.$ $n_{i,j}$ is the stoichiometric coefficient of compound j in reaction i; $\Delta_f \mathrm{G}'^{o}_j$ is the standard Gibbs free energy of formation of compound j; x_j is the concentration of the compound j; R is the ideal gas constant, $R = 8.31 \cdot 10^{-3}\ \frac{KJ}{K\ mol}$, and T Is the temperature. In this case, $T = 298\ K$.

The value of the Gibbs free energy determines the directionality of the corresponding reaction and the thermodynamically feasible pathways. With this formulation, we included the concentrations of the metabolites as variables in the mathematical formulation. TFA allows the integration of metabolomics data into the model.

2.4.4 iMM: characterizing the extracellular *in silico* minimal media

iMM is formulated as a MILP problem that introduces new variables and constraints to the TFA problem to find the minimum set of extracellular metabolites necessary to simulate growth or a specific metabolic task with the GEM [37, 38]. iMM identifies the minimum number of boundary reactions (uptakes and secretions) that need to be active. The method defines new binary variables in the TFA problem that represent the state of each boundary reaction, active or inactive. New constraints link the new binary variables to the corresponding reaction rates such that if the reaction is inactive, then it should not carry flux. The objective of the problem is to maximize the number of inactive reactions.

Assuming a network with m metabolites and n reactions, the mathematical formulation of the iMM problem is the following:

$$
\begin{aligned}
&objective\ function && \max \sum_{k=1}^{n_b} \boldsymbol{z}_k \\
& && \text{subject to} \\
&FBA\ constraints && \boldsymbol{S} \cdot \boldsymbol{v} = \boldsymbol{0}, \\
& && \boldsymbol{v}_L \leq \boldsymbol{v} \leq \boldsymbol{v}_U, \\
&TFA\ constraints && \Delta_r \mathrm{G}'_i = \sum_{j=1}^{m} n_{i,j} \Delta_f \mathrm{G}'^{o}_j + RT \ln\left(\prod_{j=1}^{m} x_j^{n_{i,j}}\right), i = 1, \dots, n, \\
& && \Delta_r \mathrm{G}'_i - M + M \cdot b_i^F \leq 0 \\
& && -\Delta_r \mathrm{G}'_i - M + M \cdot b_i^R \leq 0 \\
& && v_i^{F,R} - M \cdot b_i^{F,R} \leq 0 \\
& && b_i^F + b_i^R \leq 1 \\
&IMM\ constraints && \boldsymbol{b}^F + \boldsymbol{b}^R + C \cdot \boldsymbol{z} \leq C,
\end{aligned}
$$

where n_b is the total number of boundary reactions in the model, $\boldsymbol{z}_{rxn}$ are new binary variables for all the boundary reactions, $\boldsymbol{S}$ is the stoichiometric matrix, $\boldsymbol{v}$ are the net fluxes for all the reactions and v_i^F, v_i^R are the corresponding net-forward and net-reverse fluxes, so that, $v_i = v_i^F - v_i^R$, $for\ all\ i = 1, \dots, n$. $\boldsymbol{v}_L$ and $\boldsymbol{v}_U$ are the lower and upper bound, respectively, for all the reactions in the network. $\Delta_r \mathbf{G}'$ is the Gibb's free energy of the reactions defined in TFA. $\boldsymbol{b}^F$ and $\boldsymbol{b}^R$ are the binary variables for the forward or reverse fluxes of all the reactions (coupled to TFA). M is a big constant (bigger than all upper bounds) and C is an arbitrary large number. In this case, if $z_{rxn,i} = 0$, then reaction i is active.

2.4.5 redGEM, redGEMX, and lumpGEM: reducing human GEMs

The redGEM, redGEMX, and lumpGEM algorithms seek to generate systematic reductions of the GEMs starting from chosen subsystems (or lists of reactions and metabolites, such as the synthesis pathway of a target metabolite), based on the studied physiology and the specific parts of the metabolism that are of interest.

redGEM

redGEM is a published algorithm [35] that extracts the reactions that pairwise-connect the initial subsystems from the GEM, generating a connected network named the core network.

The inputs for redGEM are (i) the GEM, (ii) the starting subsystems or an initial set of reactions, (iii) the extracellular medium metabolites, (iv) a list with the GEM cofactor pairs, and (v) the desired degree of connectivity. The algorithm then performs an expansion (by graph search) of the starting subsystems by finding the reactions that pairwise-connect the subsystems up to the selected degree (see [35] for further details). For example, for a degree equal to 2, it will connect the metabolites from the starting subsystem that are one and two reactions away in the GEM.

redGEMX

redGEMX is a newly formulated algorithm that finds the pathways in the GEM that connect the extracellular medium to the core network generated with redGEM (Figure 2.6). These pathways are added to the core network.

The redGEMX method involves five steps:

1. Classify the extracellular metabolites of the GEM into 3 classes:

    (a) those that are part of the medium that we want to connect,

    (b) those that are already present in the inter-connected subsystems network,

    (c) those that do not belong to (a) nor (b).

2. Classify the reactions from the GEM into 2 classes:

    (a) those that belong to the inter-connected subsystems network (core-reactions),

(b) those that do not belong to the inter-connected subsystems network (non-core reactions).

3. Block the flux through the reactions in the GEM that involve only extracellular metabolites.

4. Block the flux through the boundary reactions of other metabolites in the GEM (1c). Steps 3 and 4 guarantee that the subnetwork reaches the core network.

5. Force the uptake of a medium metabolite (1a, one-by-one) and minimize the number of non-core reactions (2b) required to connect this extracellular metabolite to any core metabolite participating in a core reaction (2a). Note that the subnetwork will contain any reaction required to balance the by-products secreted by the subnetwork and/or the core network.

The redGEMX is a MILP problem that is formulated as follows:

i. Consider the TFA problem of the model that we want to reduce.
ii. Create binary variables z_i for each non-core reaction (2b). Non-core reactions are denoted as R^{nc}.
iii. Generate a constraint that controls the flux for each non-core reaction:

$$\boldsymbol{b}^F + \boldsymbol{b}^R + \boldsymbol{z} \leq \mathbf{1},$$

where $\boldsymbol{b}^F$ and $\boldsymbol{b}^R$ are the binary variables for the forward and reverse fluxes of all the reactions (coupled to the TFA constraints; when $z_i = 1$, the corresponding reaction is inactive.

iv. Build the following MILP problem for each extracellular medium metabolite (1a)

$$\max \sum_{i=1}^{R^{nc}} z_{rxn,i}$$

subject to:

$$\boldsymbol{b}^F + \boldsymbol{b}^R + \boldsymbol{z} \leq 1,$$

$$v_{eM,j} \geq c,$$

where $v_{eM,j}$ is the flux of the jth extracellular medium metabolite (1a), and c is a small number.

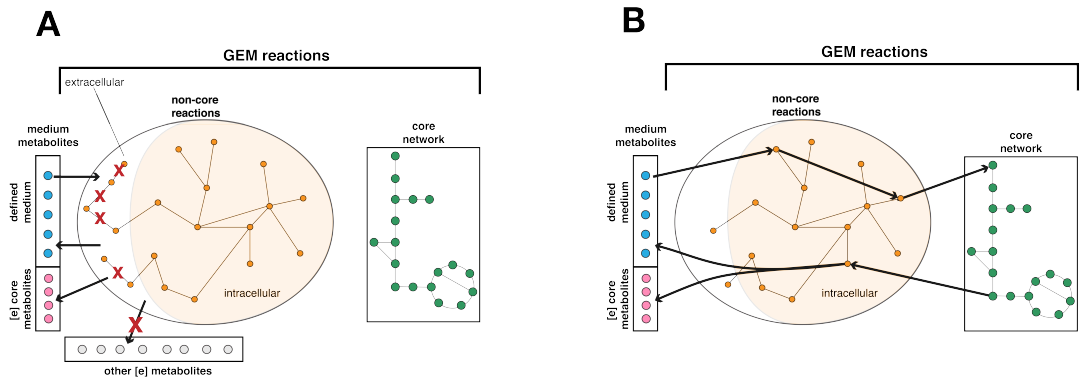


Figure 2.6. **redGEMX method**. (A) Classification of the reactions from the GEM into core (green) and non-core reactions (orange), and classification of the extracellular metabolites from the GEM into those that are part of the medium that we want to connect (blue), those that are present in the core (pink), and the others (grey). The algorithm will block the non-core reactions that involve only extracellular metabolites as well as the boundary and transport reactions of the metabolites that are not part of the medium (grey). (B) The algorithm finds the minimal set of reactions that are required to connect each of the medium metabolites (blue) to the core network, uses the core network to balance the reactions, and secretes metabolites from the medium (blue or pink).

lumpGEM

lumpGEM is a published algorithm [36] that generates elementally balanced lumped reactions for the synthesis of the biomass building blocks (BBBs). Using a MILP formulation, lumpGEM identifies the smallest subnetwork (minimum number of reactions from the GEM) required to produce each BBB from metabolites that belong to the core network using reactions from the GEM that are not part of the core. With this formulation, we can identify all the alternative subnetworks (of minimal size or larger) for the synthesis of each BBB (one by one). lumpGEM generates, for each BBB, an overall lumped reaction by adding all the reactions that constitute each subnetwork (see [36] for further details). Note here, different subnetworks can give rise to the same overall lumped reaction. This implies that although we produce all the alternative subnetworks with their associated lumped reactions, only the unique lumped reactions will be added to the final reduction.

2.4.6 Software

The simulations of this article have been done with Matlab 2017b and CPLEX 12.7.1. Escher [73] has been used to draw the subnetworks in the figures.

2.4.7 Data and code availability

The data, models and the scripts to generate the results and perform the postprocessing for this paper are available at https://github.com/EPFL-LCSB/redhuman.

The code for TFA is available at https://github.com/EPFL-LCSB/mattfa. The code to reduce the human GEMs (redGEM), to connect the extracellular medium to the core (redGEMX) and to generate the biosynthetic lumped reactions (lumpGEM) are available at https://github.com/EPFL-LCSB /redgem.

2.5 Author contribution

The work of this chapter is published in *Nature communications* with the following reference:

Masid, M., Ataman, M. & Hatzimanikatis, V. Analysis of human metabolism by reducing the complexity of the genome-scale models using redHUMAN. Nat Commun 11, 2821 (2020). https://doi.org/10.1038/s41467-020-16549-2

For this work, the collection of thermodynamic data for human cells, the design and coding of the algorithm redGEMX, the application of the methods redGEM and lumpGEM to the human GEMs, and the review of the manuscript were performed by Dr. Meric Ataman and Maria Masid under the supervision of Prof. Vassily Hatzimanikatis. The thermodynamic curation of the human GEMs, the reduction of the human GEMs with redHUMAN, the analysis of the subnetworks, the collection of leukemia data from literature, the integration of the omics data into the reduced models, the study of metabolism in leukemia cells, as well as, the writing of the manuscript and all the figures in the manuscript were performed by Maria Masid under the supervision of Prof. Vassily Hatzimanikatis.

References

1. Hanahan, D. and R.A. Weinberg, *Hallmarks of Cancer: The Next Generation.* Cell, 2011. 144(5): p. 646-674.
2. Mardinoglul, A. and J. Nielsen, New paradigms for metabolic modeling of human cells. Current Opinion in Biotechnology, 2015. 34: p. 91-97.
3. Orth, J.D., et al., A comprehensive genome-scale reconstruction of Escherichia coli metabolism-2011. Molecular Systems Biology, 2011. 7.
4. Kerkhoven, E.J., Pomraning, K. R., Baker, S. E. and Nielsen, J., Regulation of amino-acid metabolism controls flux to lipid accumulation in Yarrowia lipolytica. npj Systems Biology and Applications, 2016. 2(16005).
5. Heavner, B.D., et al., Version 6 of the consensus yeast metabolic network refines biochemical coverage and improves model performance. Database-the Journal of Biological Databases and Curation, 2013.
6. Mardinoglu, A., F. Gatto, and J. Nielsen, Genome-scale modeling of human metabolism a systems biology approach. Biotechnology Journal, 2013. 8(9): p. 985.
7. O'Brien, E.J., J.M. Monk, and B.O. Palsson, Using Genome-scale Models to Predict Biological Capabilities. Cell, 2015. 161(5): p. 971-987.
8. Orth, J.D., I. Thiele, and B.O. Palsson, What is flux balance analysis? Nature Biotechnology, 2010. 28(3): p. 245-248.
9. Henry, C.S., L.J. Broadbelt, and V. Hatzimanikatis, Thermodynamics-based metabolic flux analysis. Biophysical Journal, 2007. 92(5): p. 1792-1805.
10. Soh, K.C. and V. Hatzimanikatis, Network thermodynamics in the post-genomic era. Current Opinion in Microbiology, 2010. 13(3): p. 350-357.
11. Ataman, M. and V. Hatzimanikatis, Heading in the right direction: thermodynamics-based network analysis and pathway engineering. Current Opinion in Biotechnology, 2015. 36: p. 176-182.
12. Soh, K.C.a.H., V., Constraining the flux space using thermodynamics and integration of metabolomics data. Methods Mol Biol (Clifton, N.J.), 2014(1191): p. pp. 49-63.

13. Salvy, P., et al., pyTFA and matTFA: a Python package and a Matlab toolbox *for Thermodynamics-based Flux Analysis.* Bioinformatics, 2018. **35**(1): p. 167-169.

14. Zhang, C., et al., *ESS: A Tool for Genome-Scale Quantification of Essentiality Score for Reaction/Genes in Constraint-Based Modeling.* Frontiers in Physiology, 2018. **9**.

15. Lewis, N.E., H. Nagarajan, and B.O. Palsson, *Constraining the metabolic genotype-phenotype relationship using a phylogeny of in silico methods.* Nature Reviews Microbiology, 2012. **10**(4): p. 291-305.

16. Nielsen, J., *Systems Biology of Metabolism.* Annual Review of Biochemistry, Vol 86, 2017. **86**: p. 245-275.

17. Bordbar, A. and B.O. Palsson, *Using the reconstructed genome-scale human metabolic network to study physiology and pathology.* Journal of Internal Medicine, 2012. **271**(2): p. 131-141.

18. Zhang, C. and Q. Hua, *Applicatons of Genome-Scale Metabolic Models in Biotechnology and Systems Medicine.* Frontiers in Physiology, 2016. **6**.

19. Chakrabarti, A., et al., *Towards kinetic modeling of genome-scale metabolic networks without sacrificing stoichiometric, thermodynamic and physiological constraints.* Biotechnol J, 2013. **8**(9): p. 1043-57.

20. Miskovic, L., et al., *Rites of passage: requirements and standards for building kinetic models of metabolic phenotypes.* Current Opinion in Biotechnology, 2015. **36**: p. 146-153.

21. Srinivasan, S., W.R. Cluett, and R. Mahadevan, *Constructing kinetic models of metabolism at genome-scales: A review.* Biotechnology Journal, 2015. **10**(9): p. 1345-1359.

22. Stanford, N.J., et al., *Systematic Construction of Kinetic Models from Genome-Scale Metabolic Networks.* Plos One, 2013. **8**(11).

23. Duarte, N.C., et al., *Global reconstruction of the human metabolic network based on genomic and bibliomic data.* Proceedings of the National Academy of Sciences of the United States of America, 2007. **104**(6): p. 1777-1782.

24. Thiele, I. and B.O. Palsson, *A protocol for generating a high-quality genome-scale metabolic reconstruction.* Nature Protocols, 2010. **5**(1): p. 93-121.

25. Mardinoglu, A., et al., *Genome-scale metabolic modelling of hepatocytes reveals serine deficiency in patients with non-alcoholic fatty liver disease.* Nature Communications, 2014. **5**.

26. Thiele, I., et al., *A community-driven global reconstruction of human metabolism.* Nature Biotechnology, 2013. **31**(5): p. 419-+.

27. Swainston, N., et al., *Recon 2.2: from reconstruction to model of human metabolism.* Metabolomics, 2016. **12**(7).

28. Brunk, E., et al., *Recon3D enables a three-dimensional view of gene variation in human metabolism.* Nature Biotechnology, 2018. **36**(3): p. 272-+.

29. Ebrahim, A., et al., *Multi-omic data integration enables discovery of hidden biological regularities.* Nature Communications, 2016. **7**.

30. Aurich, M.K., et al., *Prediction of intracellular metabolic states from extracellular metabolomic data.* Metabolomics, 2015. **11**(3): p. 603-619.

31. Schmidt, B.J., et al., *GIM(3)E: condition-specific models of cellular metabolism developed from metabolomics and expression data.* Bioinformatics, 2013. **29**(22): p. 2900-2908.

32. Becker, S.A. and B.O. Palsson, *Context-specific metabolic networks are consistent with experiments.* PLoS Comput Biol, 2008. **4**(5): p. e1000082.

33. Wang, Y., J.A. Eddy, and N.D. Price, *Reconstruction of genome-scale metabolic models for 126 human tissues using mCADRE.* BMC Syst Biol, 2012. **6**: p. 153.

34. Agren, R., et al., *Identification of anticancer drugs for hepatocellular carcinoma through personalized genome-scale metabolic modeling.* Mol Syst Biol, 2014. **10**: p. 721.

35. Ataman, M., et al., *redGEM: Systematic reduction and analysis of genome-scale metabolic reconstructions for development of consistent core metabolic models.* Plos Computational Biology, 2017. **13**(7).

36. Ataman, M. and V. Hatzimanikatis, *lumpGEM: Systematic generation of subnetworks and elementally balanced lumped reactions for the biosynthesis of target metabolites.* Plos Computational Biology, 2017. **13**(7).

37. Tymoshenko, S., et al., *Metabolic Needs and Capabilities of Toxoplasma gondii through Combined Computational and Experimental Analysis.* PLoS Comput Biol, 2015. **11**(5): p. e1004261.

38. Chiappino-Pepe, A., et al., *Bioenergetics-based modeling of Plasmodium falciparum metabolism reveals its essential genes, nutritional requirements, and thermodynamic bottlenecks.* Plos Computational Biology, 2017. **13**(3).

39. Folger, O., et al., *Predicting selective drug targets in cancer through metabolic networks.* Molecular Systems Biology, 2011. **7**.

40. Zielinski, D.C., et al., *Systems biology analysis of drivers underlying hallmarks of cancer cell metabolism.* Scientific Reports, 2017. **7**.

41. Shlomi, T., et al., *Genome-Scale Metabolic Modeling Elucidates the Role of Proliferative Adaptation in Causing the Warburg Effect.* Plos Computational Biology, 2011. **7**(3).

42. Warburg, O., *On the metabolism of cancer cells.* Naturwissenschaften, 1924. **12**: p. 1131-1137.

43. Fenninger, L.D. and G.B. Mider, *Energy and Nitrogen Metabolism in Cancer.* Advances in Cancer Research, 1954. **2**: p. 229-251.

44. Cairns, R.A., I.S. Harris, and T.W. Mak, *Regulation of cancer cell metabolism.* Nature Reviews Cancer, 2011. **11**(2): p. 85-95.

45. Dang, C.V., *Links between metabolism and cancer.* Genes & Development, 2012. **26**(9): p. 877-890.

46. Di Filippo, M., et al., *Zooming-in on cancer metabolic rewiring with tissue specific constraint-based models.* Computational Biology and Chemistry, 2016. **62**: p. 60-69.

47. Hosios, A.M., et al., *Amino Acids Rather than Glucose Account for the Majority of Cell Mass in Proliferating Mammalian Cells.* Developmental Cell, 2016. **36**(5): p. 540-549.

48. Hay, N., *Reprogramming glucose metabolism in cancer: can it be exploited for cancer therapy?* Nature Reviews Cancer, 2016. **16**(10): p. 635-649.

49. Wise, D.R. and C.B. Thompson, *Glutamine addiction: a new therapeutic target in cancer.* Trends in Biochemical Sciences, 2010. **35**(8): p. 427-433.

50. Kanarek, N., et al., *Histidine catabolism is a major determinant of methotrexate sensitivity.* Nature, 2018. **559**(7715): p. 632-636.

51. Henry, C.S., L.J. Broadbelt, and V. Hatzimanikatis, *Discovery and analysis of novel metabolic pathways for the biosynthesis of industrial chemicals: 3-hydroxypropanoate.* Biotechnol Bioeng, 2010. **106**(3): p. 462-73.

52. Du, B., et al., *Thermodynamic favorability and pathway yield as evolutionary tradeoffs in biosynthetic pathway choice.* Proc Natl Acad Sci U S A, 2018. **115**(44): p. 11339-11344.

53. Feist, A.M. and B.O. Palsson, *The biomass objective function.* Curr Opin Microbiol, 2010. **13**(3): p. 344-9.

54. Harayama, T. and H. Riezman, *Understanding the diversity of membrane lipid composition.* Nature Reviews Molecular Cell Biology, 2018. **19**(5): p. 281-296.

55. Kanehisa, M., et al., *KEGG: new perspectives on genomes, pathways, diseases and drugs.* Nucleic Acids Research, 2017. **45**(D1): p. D353-D361.

56. Jain, M., et al., *Metabolite Profiling Identifies a Key Role for Glycine in Rapid Cancer Cell Proliferation.* Science, 2012. **336**(6084): p. 1040-1044.

57. Kilburn, D.G., M.D. Lilly, and F.C. Webb, *The energetics of mammalian cell growth.* J Cell Sci, 1969. **4**(3): p. 645-54.

58. DeBerardinis, R.J. and N.S. Chandel, *Fundamentals of cancer metabolism.* Sci Adv, 2016. **2**(5): p. e1600200.

59. Hatzimanikatis, V., et al., *Exploring the diversity of complex metabolic networks.* Bioinformatics, 2005. **21**(8): p. 1603-9.

60. Jankowski, M.D., et al., *Group contribution method for thermodynamic analysis of complex metabolic networks.* Biophysical Journal, 2008. **95**(3): p. 1487-1499.

61. Wilhelm, M., et al., *Mass-spectrometry-based draft of the human proteome.* Nature, 2014. **509**(7502): p. 582-7.

62. Uhlen, M., et al., *Proteomics. Tissue-based map of the human proteome.* Science, 2015. **347**(6220): p. 1260419.

63. Jerby, L., T. Shlomi, and E. Ruppin, *Computational reconstruction of tissue-specific metabolic models: application to human liver metabolism.* Molecular Systems Biology, 2010. **6**.

64. Bordbar, A., et al., *Elucidating dynamic metabolic physiology through network integration of quantitative time-course metabolomics.* Scientific Reports, 2017. **7**.

65. Sanchez, B.J., et al., *Improving the phenotype predictions of a yeast genome-scale metabolic model by incorporating enzymatic constraints.* Molecular Systems Biology, 2017. **13**(8).

66. Salvy, P. and V. Hatzimanikatis, *ETFL: A formulation for flux balance models accounting for expression, thermodynamics, and resource allocation constraints.* bioRxiv doi: 10.1101/590992, 2019.

67. Pandey, V., et al., *TEX-FBA: A constraint-based method for integrating gene expression, thermodynamics, and metabolomics data into genome-scale metabolic models.* bioRxiv doi: 10.1101/536235, 2019.

68. Yizhak, K., et al., *Integrating quantitative proteomics and metabolomics with a genome-scale metabolic network model.* Bioinformatics, 2010. **26**(12): p. i255-i260.

69. Moretti, S., et al., *MetaNetX/MNXref - reconciliation of metabolites and biochemical reactions to bring together genome-scale metabolic networks.* Nucleic Acids Research, 2016. **44**(D1): p. D523-D526.

70. Overbeek, R., et al., *The subsystems approach to genome annotation and its use in the project to annotate 1000 genomes.* Nucleic Acids Research, 2005. **33**(17): p. 5691-5702.

71. Hastings, J., et al., *ChEBI in 2016: Improved services and an expanding collection of metabolites.* Nucleic Acids Research, 2016. **44**(D1): p. D1214-D1219.

72. Wishart, D.S., et al., *HMDB 4.0: the human metabolome database for 2018.* Nucleic Acids Research, 2018. **46**(D1): p. D608-D617.

73. King, Z.A., et al., *Escher: A Web Application for Building, Sharing, and Embedding Data-Rich Visualizations of Biological Pathways.* PLoS Comput Biol, 2015. **11**(8): p. e1004321.

Appendix A

Curation of the generic lipids in human GEMs

Note S2.1: We curated the synthesis of the generic fatty acids in both models. Some compounds from the lipid metabolism of the human GEMs contain R-groups as generic compounds representing fatty acids with different chain lengths. The biosynthesis of fatty acids starts with the conversion of acetyl-coA into malonyl-coA. Next, the primary fatty acid synthesized is palmitate (C16:0) that can later enter the elongation cycle to produce fatty acids with longer chains. We connected palmitate to the R-groups using reactions already defined in the original Recon 2 and Recon 3D models. These reactions substitute the elongation process and define directly the generic fatty acids (R-groups).

redGEM: network expansion

Table S2.1: redGEM statistics. Expansion of the starting subsystems by pairwise connections for different degrees. D=0 does not include interconnections across subsystems, only the intra-expansion of the initial subsystems

	Recon 2				**Recon 3D**			
Degree of connection	D = 0	D = 1	D = 2	D = 3	D = 0	D = 1	D = 2	D = 3
Number of metabolites	254	**356**	434	487	300	**440**	530	592
Number of reactions	346	**617**	1286	1412	416	**796**	1451	1620

redGEMX: extracellular metabolites connections to core

Table S2.2: **redGEMX connections.** Size of subnetworks [and number of alternatives] to connect the medium metabolites to the core subsystems. As an example, the two last columns show the core metabolites and the subsystems to which the extracellular connects for one alternative.

		Extracellular metabolites	Recon 2 Num. of reactions [num of alternatives]	Recon 3D Num. of reactions [num of alternatives]
Amino acids	Glutamate Family	L-glutamate	10 [8]	9 [1]
		L-glutamine	11 [3]	11 [4]
		L-proline	13 [1]	12 [1]
		L-arginine	11 [3]	8 [1]
	Serine Family	L-serine	6 [1]	6 [5]
		glycine	18 [1]	8 [21]
	Pyruvate Family	L-alanine	12 [27]	9 [6]
		L-valine	24 [57]	23 [3]
		L-isoleucine	27 [3]	23 [2]
		L-leucine	32 [1]	23 [1]
	Aspartate Family	L-aspartate	13 [1]	7 [4]
		L-asparagine	14 [2]	9 [1]
		L-lysine	33 [13]	40 [33]
		L-methionine	42 [1]	22 [4]
		L-threonine	21 [1]	6 [1]
	Aromatic Family	L-tryptophan	39 [8]	38 [1]
		L-tyrosine	27 [3]	25 [1]
		L-phenylalanine	27 [2]	27 [1]
		L-histidine	22 [3]	20 [2]
Central carbon compounds		D-glucose	5 [1]	5 [1]
		L-lactate	5 [1]	5 [1]
		L-Malate	13 [2]	9 [5]
		Citrate	10 [2]	5 [9]
Vitamin		Choline	14 [10]	11 [7]
Urea cycle products		Ornithine	7 [3]	7 [3]
		Urea	7 [3]	7 [3]

lumpGEM: subnetworks of lumped reactions

Table S2.3: **Lumped reactions for Recon 2 and Recon 3D.** Size of lumped reactions for Recon 2 and Recon 3D and corresponding number of alternatives. The rest of the BBBs either are directly uptaken (available in the medium) or they can be produced in the core network.

	Biomass building blocks	Recon 2		Recon 3D	
		Network size	Number alternatives	Network size	Number alternatives
Amino acid	L-Cysteine	Produced by the core	Produced by the core	2	1
Nucleotides	ATP	10	1	11	1
	GTP	9	2	10	4
	CTP	5	1	5	1
	UTP	5	1	5	1
	dGTP	11	4	12	8
	dCTP	7	1	7	1
	dATP	12	1	13	2
	dTTP	11	4	11	6
Lipid	Cholesterol	21	64	21	144
Phospholipids & Sphingolipids	1-Phosphatidyl-1D-Myo-Inositol	10	1	10	4
	Phosphatidylserine	8	1	8	4
	Phosphatidylcholine	9	1	8	4
	Phosphatidylethanolamine	9	1	8	4
	Phosphatidylglycerol	11	7	11	14
	Cardiolipin	12	7	12	14
	Sphingomyelin	12	2	11	6
	Growth Associated Maintenance	10	1	11	2

Table S2.4: Lumped reactions for phosphatidyl serine in Recon 3D. Stoichiometry of the alternative lumped reactions for phosphatidylserine in Recon 3D

	Lumped reactions for phosphatidylserine in Recon 3D			
	Lumped Reaction 1	Lumped Reaction 2	Lumped Reaction 3	Lumped Reaction 4
ATP	-16	-16	-8	-8
H^+	-53	-53	-25	-25
NADPH	-28	-28	-14	-14
HCO3	-14	-14	-7	-7
ACCOA	-16	-16	-8	-8
SER_L	-1	0	-1	0
CHOL	0	-1	0	-1
LNLCCOA	0	0	-1	-1
GLYC3P	-1	-1	-1	-1
H2O	13	13	7	7
ADP	14	14	7	7
Pi	14	14	7	7
NADP	28	28	14	14
CO2	14	14	7	7
COA	16	16	9	9
AMP	2	2	1	1
PPi	2	2	1	1
PS	1	0	1	0
PCHOL	0	1	0	1
	53 H^+ + 16 ATP + 28 NADPH + 14 HCO3 + 16 ACCOA + SER_L + GLYC3P -> 13 H2O + 14 ADP + 14 Pi + 28 NADP + 14 CO2 + 16 COA + 2 AMP + 2 PPi + PS	53 H^+ + 16 ATP + 28 NADPH + 14 HCO3 + 16 ACCOA + CHOL + GLYC3P -> 13 H2O + 14 ADP + 14 Pi + 28 NADP + 14 CO2 + 16 COA + 2 AMP + 2 PPi + PCHOL	25 H^+ + 8 ATP + 14 NADPH + 7 HCO3 + 8 ACCOA + SER_L + LNLCCOA + GLYC3P -> 7 H2O + 7 ADP + 7 Pi + 14 NADP + 7 CO2 + 9 COA + AMP + PPi + PS	25 H^+ + 8 ATP + 14 NADPH + 7 HCO3 + 8 ACCOA + LNLCCOA + CHOL + GLYC3P -> 7 H2O + 7 ADP + 7 Pi + 14 NADP + 7 CO2 + 9 COA + AMP + PPi + PCHOL

Leukemia physiology: extracellular data integrated in the models

Table S2.5: **Leukemia physiology**. Data used to constrain the intake and secretion of extracellular metabolites.

	Reaction in model	Lower bound	Upper bound
Citrate	EX_cit_e	-0.0007	0.0015
Malate	EX_mal_L_e	0.0002	0.0012
Choline	EX_chol_e	-0.0078	-0.0004
Ornithine	EX_orn_e	0.0077	0.0498
Alanine	EX_ala_L_e	-0.0137	0.1695
Arginine	EX_arg_L_e	-0.0752	-0.0090
Asparagine	EX_asn_L_e	-0.0151	-0.0035
Aspartate	EX_asp_L_e	-0.0098	0.0051
Glucose	EX_glc_e	-3.3230	-0.4623
Glutamate	EX_glu_L_e	0.0089	0.0883
Glutamine	EX_gln_L_e	-0.6398	-0.1672
Glycine	EX_gly_e	-0.0031	0.0139
Isoleucine	EX_ile_L_e	-0.0420	-0.0083
Lactate	EX_lac_L_e	0.5531	3.7791
Leucine	EX_leu_L_e	-0.0508	-0.0099
Lysine	EX_lys_L_e	-0.0697	-0.0170
Phenylalanine	EX_phe_L_e	-0.0206	-0.0045
Proline	EX_pro_L_e	-0.0015	0.0145
Serine	EX_ser_L_e	-0.1140	-0.0262
Threonine	EX_thr_L_e	-0.0451	-0.0092
Tryptophan	EX_trp_L_e	-0.0054	-0.0008
Tyrosine	EX_tyr_L_e	-0.0276	-0.0054
Valine	EX_val_L_e	-0.0483	-0.0098
Methionine	EX_met_L_e	-0.0186	-0.0041
Oxygen	EX_o2_e	-2	0
Growth	biomass	0	0.0354
ATP maintenance	ATPM	1.07	100

Table S2.6: **Leukemia physiology**. Data used to constrain the extracellular concentrations of the metabolites.

	Variable in model	Lower bound	Upper bound
3-Hydroxyanthranilate	LC_3hanthrn_e	-17.6225	-17.5215
4-Aminobutanoate	LC_4abut_e	-15.6566	-12.998
Acetoacetate	LC_acac_e	-14.0479	-13.567
Adenosine	LC_adn_e	-100	-16.5006
S-Adenosyl-L-Homocysteine	LC_ahcys_e	-17.1933	-16.2063
2-Oxoglutarate	LC_akg_e	-15.6860	-15.1312
L-Alanine	LC_ala_L_e	-10.9925	-7.6359
L-Arginine	LC_arg_L_e	-6.9545	-6.8303
L-Asparagine	LC_asn_L_e	-8.0950	-7.9203
L-Aspartate	LC_asp_L_e	-8.8353	-8.6895
Choline	LC_chol_e	-10.8158	-10.5017
Citrate	LC_cit_e	-12.2716	-11.4708
Creatine	LC_creat_e	-10.8187	-10.6980
L-Carnitine	LC_crn_e	-14.1057	-13.8208
Fumarate	LC_fum_e	-13.4448	-13.1200
L-Glutamine	LC_gln_L_e	-7.6293	-6.5510
L-Glutamate	LC_glu_L_e	-8.6402	-8.1777
Glycine	LC_gly_e	-8.8527	-8.6086
Guanidinoacetic Acid	LC_gudac_e	-9.6591	-9.1620
L-Homocysteine	LC_hcys_L_e	-15.9947	-15.0187
3-Hydroxy-L-Kynurenine	LC_hLkynr_e	-19.5847	-18.4311
Isocitric Acid	LC_icit_e	-15.1765	-14.8061
L-Isoleucine	LC_ile_L_e	-8.1052	-7.9730
L-2-Aminoadipate	LC_L2aadp_e	-15.9991	-15.4878
(S)-Lactate	LC_lac_L_e	-5.0995	-4.6995
L-Kynurenine	LC_Lkynr_e	-15.3602	-14.7654
L-Lysine	LC_lys_L_e	-9.1042	-8.4909
L-Homoserine	LC_hom_L_e	-10.6130	-10.2175
(S)-Malate	LC_mal_L_e	-12.2526	-11.9394
L-Methionine	LC_met_L_e	-9.9144	-9.3876
Ornithine	LC_orn_e	-8.9184	-8.5935
L-Phenylalanine	LC_phe_L_e	-10.0679	-9.5613
L-Serine	LC_ser_L_e	-10.6420	-8.7444
Spermidine	LC_spmd_e	-15.2091	-14.8503
Succinate	LC_succ_e	-11.2943	-11.1637
Taurine	LC_taur_e	-13.1674	-12.0361
L-Threonine	LC_thr_L_e	-9.45495	-8.9928
L-Tryptophan	LC_trp_L_e	-11.2214	-10.7397
L-Tyrosine	LC_tyr_L_e	-9.2399	-8.9257
L-Valine	LC_val_L_e	-9.4606	-9.0726

Gene essentiality

Table S2.7: **Gene essentiality analysis.** List of genes that are essential in the reduced Recon 2 models, and the corresponding reactions associated to these genes. The third column indicates if the gene is essential in the GEM.

GENE	REACTIONS	Essential in GEM
1738.1	GCC2am I GCC2bim I GCC2cm I GCCam I GCCbim I GCCcm I PDHm I r1154	-
2531.1	3DSPHR	Yes
2194.1	KAS8	Yes
471.1	AICART I IMPC	Yes
10606.1	AIRCr I PRASCS	Yes
790.1	CBPS I ASPCTr I DHORTS	Yes
7108.1	C14STRr I r0780	Yes
50814.1	C3STDH1Pr I C4STMO2Pr I C3STDH1r	Yes
3295.1	C3STKR2r	Yes
6307.1	C4STMO1r	Yes
54675.1	CLS_hs	Yes
875.1	CYSTS	Yes
51727.1	UMPK I UMPK2 I UMPK3 I UMPK4 I UMPK5 I UMPK7 I CYTK8 I CYTK6 I CYTK7 I CYTK1 I CYTK10	Yes
2987.1	GK1	-
1718.1	r0783 I DSREDUCr I r1380	-
1717.1	DHCR72r I DHCR71r	Yes
1719.1	DHFR I r0224	Yes
1723.1	DHORD9	Yes
9453.1	DMATT I GRTT	-
4597.1	DPMVDc	Yes
1841.1	DTMPK I NDP8	-
10682.1	EBP1r I EBP2r I r1381	Yes
2819.1	G3PD1 I r0202	-
2618.1	r0666	Yes
5471.1	GLUPRT	Yes
3156.1	r0488 I HMGCOARc	Yes
4047.1	LNSTLSr	Yes
4598.1	MEVK1c	Yes
51477.1	MI1PS	Yes
7372.1	OMPDC I ORPT	Yes
5338.1	RE3273C I RE3301C	-
114971.1	PGPP_hs	Yes
9489.1	PGPPT	Yes
10654.1	PMEVKc	Yes
5198.1	PRFGS	Yes
23761.1	PSDm_hs	-
9791.1	PSSA1_hs	-
6240.1	r0472 I r0474 I r0475	-
6241.1	r0472 I r0474 I r0475	-
50484.1	r0472 I r0474 I r0475	-
22934.1	RPI I r0249	Yes
10558.1	SERPT	Yes
9517.1	SERPT	Yes
55304.1	SERPT	Yes
259230.1	SMS	Yes
6713.1	SQLEr	Yes
2222.1	SQLSr	Yes
7298.1	TMDS	Yes
1595.1	r0781	Yes

Table S2.8: **Gene essentiality analysis.** List of genes that are essential in the in the reduced Recon 3D models, and the corresponding reactions associated to these genes. The third column indicates if the gene is essential in the GEM.

GENE	REACTIONS	Essential in GEM
1738.1	GCC2am | GCC2bim | GCC2cm | GCCam | GCCbim | GCCcm | PDHm | r1154	-
8050.1	2OXOADOXm | AKGDm | PDHm	-
1743.1	2OXOADOXm | AKGDm	-
2531.1	3DSPHR	Yes
10606.1	AIRCr | PRASCS	-
790.1	CBPS | ASPCTr | DHORTS	Yes
7108.1	C14STRr	-
50814.1	C3STDH1Pr | C4STMO2Pr | C3STDH1r	Yes
54675.1	CLS_hs	Yes
51727.1	UMPK2 | UMPK3 | UMPK4 | UMPK5 | UMPK7 | UMPK | CYTK1 | CYTK10 | CYTK6 | CYTK8 | CYTK7	-
1717.1	DHCR72r | DHCR71r	-
1723.1	DHORD9	Yes
4597.1	DPMVDc | DPMVDx	Yes
10682.1	EBP1r | EBP2r	Yes
2194.1	KAS8	-
1719.1	DHFR | r0224	Yes
2618.1	r0666	-
3158.1	HMGCOASim	-
4047.1	LNSTLSr	Yes
4598.1	MEVK1c | MEVK1x	Yes
51477.1	MI1PS | HMR_6572	Yes
1841.1	DTMPK | NDP8	-
7372.1	ORPT | OMPDC	Yes
114971.1	PGPP_hs	-
9489.1	PGPPT	Yes
10654.1	PMEVKc | PMEVKx	Yes
10558.1	SERPT	Yes
9517.1	SERPT	Yes
55304.1	SERPT	Yes
259230.1	SMS	-
6713.1	SQLEr	Yes
2222.1	SQLSr	Yes
2819.1	r0202 | G3PD1 | HMR_0478	Yes
875.1	CYSTS	Yes
22934.1	r0249 | RPI	-
50484.1	r0472 | r0474 | r0475	-
6240.1	r0472 | r0474 | r0475	-
6241.1	r0472 | r0474 | r0475	-
3156.1	r0488 | HMR_4630	Yes
1595.1	r0781	-
2987.1	GK1	-
471.1	AICART | IMPC	-
5471.1	GLUPRT	-
5198.1	PRFGS	-
7298.1	TMDS	Yes
7384.1	CYOR_u10mi | CYOOm2i	-
7388.1	CYOR_u10mi | CYOOm2i	-
4519.1	CYOR_u10mi | CYOOm2i	-
10975.1	CYOR_u10mi | CYOOm2i	-
7385.1	CYOR_u10mi | CYOOm2i	-
7386.1	CYOR_u10mi | CYOOm2i	-
1537.1	CYOR_u10mi | CYOOm2i	-

27089.1	CYOR_u10mi l CYOOm2i	-
7381.1	CYOR_u10mi l CYOOm2i	-
1351.1	CYOOm3i l CYOOm2i	-
1347.1	CYOOm3i l CYOOm2i	-
1329.1	CYOOm3i l CYOOm2i	-
1327.1	CYOOm3i l CYOOm2i	-
341947.1	CYOOm3i l CYOOm2i	-
1350.1	CYOOm3i l CYOOm2i	-
1349.1	CYOOm3i l CYOOm2i	-
1339.1	CYOOm3i l CYOOm2i	-
1345.1	CYOOm3i l CYOOm2i	-
9377.1	CYOOm3i l CYOOm2i	-
170712.1	CYOOm3i l CYOOm2i	-
1340.1	CYOOm3i l CYOOm2i	-
1337.1	CYOOm3i l CYOOm2i	-

Reaction-Gene expression

Table S2.9: **Reactions with gene expression**. Number of reactions associated to expressed genes in the corresponding NCI60 leukemia cell lines.

MODEL	Reduced Recon2	Reduced Recon2 Smin	Reduced Recon3	Reduced Recon3 Smin
Total number of reactions	1429	1451	1691	1738
Number of reactions with GPRs	1194	1215	1282	1317
Number of leukemia NCI60 expressed reactions	1190	1211	1281	1316
% of expressed reactions w.r.t. total number of reactions	83.28%	83,46%	75.75%	75.72%
% of expressed reactions w.r.t. number of reactions with GPRs	99.66%	99.67%	99.92%	99.92%

Thermodynamic parameters used in the curation of the human GEMs

Table S2.10: **Thermodynamic parameters**. Thermodynamic properties used for the compartments in the models.

Compartment	pH	Δ Ψ [mV]	Ionic Strength [M]	Concentrations range [M]
cytosol	7.2	0	0.15	$[10^{-11} – 0.08]$
mitochondria	8	-155		
Inner-mitochondria membrane space (*)	7.2	0		
vacuole	7	0		
peroxisome	7	12		
Golgi apparatus	6.35	0		
endoplasmic reticulum	7.2	0		
nucleus	7.2	0		
lysosome	4.7	19		
extracellular	7.4	30		$[10^{-11} – 0.1]$

(*) only the Recon 3D models have this compartment

References for pH values:

- Casey et al. *Sensors and regulators of intracellular pH*. Nature Reviews, 2010
- Alberts et al. *Molecular Biology of the cell*. 4th ed New York: Garland Science, 2002

Reference for membrane potential and ionic strength:

- Haraldsdóttir et al. *Quantitative Assignment of Reaction Directionality in a Multicompartmental Human Metabolic Reconstruction*. Biophysical Journal, 2012

Chapter 3 Model-based data integration and minimal network enrichment to identify metabolic differences across cancer types

In this chapter, we apply a suite of methods to integrate data in metabolic models to build cancer-specific models and we identify the underlying metabolic functions for the genetic alterations experienced by different tumor cells. A manuscript with the content of this chapter is *in preparation* to be published.

3.1 Introduction

Cancer research to decipher and understand the cellular alterations occurring as cancer develops and progresses experienced an enormous advance in the last century [1]. However, the molecular mechanisms underlying this disease remain still unknown. During the last decades, the advances in experimental data extraction and processing have allowed to study cancer at different scales: from the molecular to the systems level; from the genetic mutations and signaling alterations to the metabolic reprogramming in cancer cells and the evasion of the immune system [2, 3]. These modifications support the survival and fast proliferation of cancer cells and their adaptation to different environmental conditions, causing resistance to existing cancer therapies [4, 5]. In particular, the metabolic alterations affect nutrient assimilation, biosynthesis of growth

precursors, bioenergetics, and redox balance, which enhance the metabolic pathways used for survival and proliferation of cancer cells in the tumor microenvironment [6, 7].

The association between metabolism and cancer has created a resurgence of interest in the fields of systems biology and metabolic modeling to analyze and understand the metabolic changes occurring in cancer cells, both with respect to their healthy counterparts and across different types of cancer. Modeling the different phenotypes of healthy and cancer cells will help to guide more effective therapies to prevent, diagnose, and treat cancer [8, 9]. The different metabolic phenotypes emerge from differences in the pathways that the cells use to synthesize the metabolites required to perform cellular functions. Genome-scale models (GEMs) are representations of the cellular metabolism of a specific organism. They are reconstructed based on the genome of the organism, and they contain gene-protein-enzyme associations and the stoichiometric relationship of the reactions and metabolites [10]. GEMs have been extensively used to understand the metabolic alterations in cancer cells [11-17].

With the ever-expanding collection of data from different sources, tissues, cells, and patients, there is an increasing development of methods to analyze and understand the data using GEMs as scaffolds. Such methods integrate data to generate context-specific metabolic models that are used to classify the differences and similarities of the metabolism at the genome-scale across samples, and they provide a remarkable derived delineation of the genotype-phenotype relationships. Some of the methods derived in the last years include iMAT [18], tINIT [19], mCADRE [20], FASTCORE [21] and MBA [22]. Although the derived context-specific models are not consistent among methods [23] and they depend on thresholds and parameters defined within each method, they have been proven to be useful in the study of metabolism in several organisms [24, 25].

These models can be further improved to include a more detailed description of the bioenergetics of the cell and to integrate quantitative metabolomics in the models. Moreover, in many cases, the large size and complexity of GEMs, together with poor data coverage, hamper the study of metabolism. In order to overcome these challenges, methods to account for the thermodynamic properties of the metabolites and reactions [26-29] and to generate reduced versions of the GEMs focusing on specific parts of metabolism [30, 31] were created.

In this work, we introduce a suite of methods and a workflow that (i) integrate exo-metabolomics and exo-fluxomics data to define the interactions of the cells with their environment, (ii) use transcriptomics data and metabolic flux balance analysis to identify the active transport reactions specific for each cell type and (iii) maximize the consistency of the metabolic reaction rates with the transcriptomics data assigning deregulation to the reaction rates based on the deregulation of their associated genes to translate how the deregulation of gene expression is reflected in the deregulation of metabolic pathways and their flux activities.

We applied this workflow to generate three cancer-type-specific models for breast cancer, colon cancer, and ovarian cancer by integrating metabolomics, fluxomics, and transcriptomics data from the NCI60 cancer cell lines [32] into a reduced version of the human GEM Recon 3D [33]. We used the developed cancer-specific metabolic models to identify metabolic differences among them, including the essentiality of metabolic genes and enzymes, and the flexibility of reaction rates, subject to the specific tumor microenvironment as it is defined by the nutrient composition and requirements of each cancer. We identified a large number of deregulations in the metabolic fluxes and pathways that could not be captured by conventional gene expression and phenotypic analysis. Furthermore, we characterized cancer metabolic phenotypes by defining metabolic tasks for each phenotype and extracting the associated subnetworks for every metabolic task in each cancer type. We next performed subnetwork enrichment analysis to explore the deregulation of the minimal networks and assign deregulation to the metabolic tasks. We identified the cancer requirements for eleven metabolic tasks that are associated to seven phenotypes in cancer cells.

3.2 Results

3.2.1 Workflow overview

The workflow that we developed integrates data from different scales to identify cancer-type specific deregulations in metabolic pathways (Figure 3.1). The Data Integration component is composed of three steps: first, we select a metabolic network model which is not specific to a tissue or cancer type, next we integrate *omics* data to define the "network topology", and to determine the "network physiology". In the first step, we use

a model based on the latest human genome-scale metabolic model, either the full version, Recon 3D, or a reduced version. In the second step, we integrate metabolomics and fluxomics to directly constrain the bounds of the corresponding metabolites and transport reactions of the model. In addition, we use the transcriptomics data qualitatively to define the network topology by selecting the transport reactions that should be present in each cell type and which are not identified by the exo-metabolomics and exo-fluxomics data directly. In the third step, the network physiology is defined using the method TEX-FBA [32], which integrates into the model transcriptomics and proteomics data by constraining the maximum number of reaction rates according to the expression profile. These three steps generate physiology-specific metabolic models that have an adequate representation of the metabolism of a particular type of cell under specific physiological conditions.

Next, we processed the context-specific models through the Pathway Deregulation component. In this second part of the workflow, we defined a set of metabolic tasks that represent the underlying phenotypes of the physiology, and we used the cancer-specific models to generate the minimal subnetworks required to satisfy each metabolic task. Then, we used the method minimal network enrichment analysis (MiNEA) [33] to perform pathway enrichment analysis and to identify the significantly deregulated minimal subnetworks and their associated tasks in each cancer type. A comparison of the findings from MiNEA allows to identify significant differences and similarities in the molecular level deregulation across different cancer physiologies.

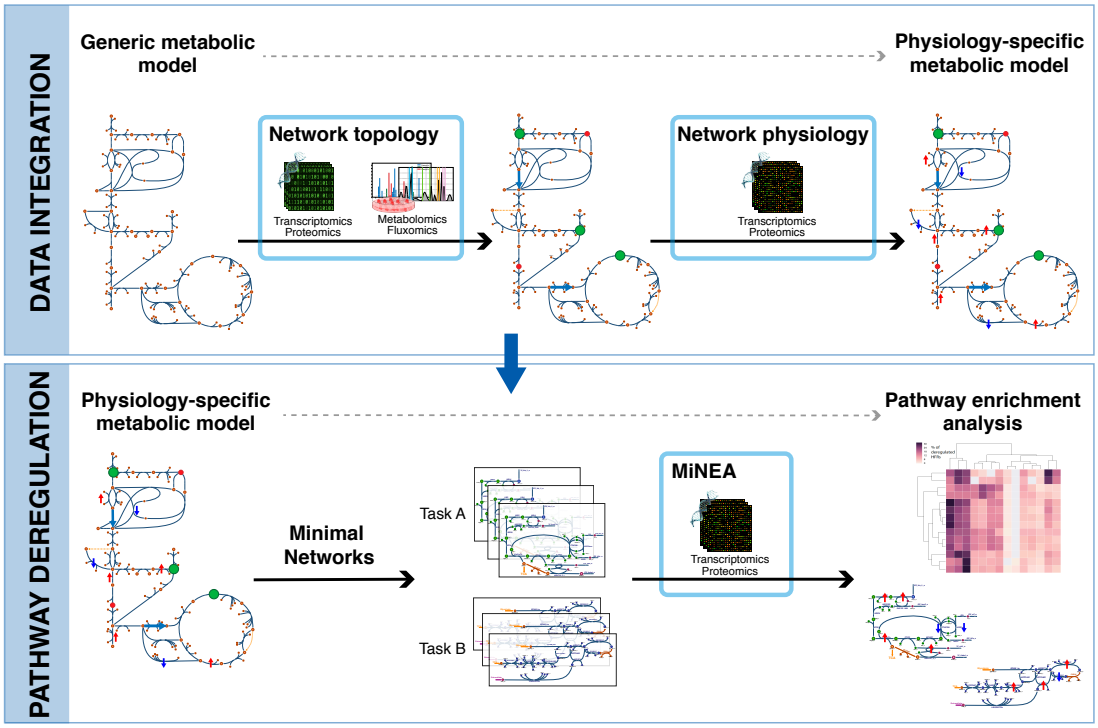


Figure 3.1. **Overview of the workflow for data integration and pathway enrichment analysis.** The workflow is divided in two parts. In the first part, *Data integration,* physiology-specific models are generated. The network topology and the network physiology are defined by integrating transcriptomics, proteomics, metabolomics, and fluxomics data into the generic metabolic model. In the second part, *Pathway deregulation,* the physiology specific models are used to generate minimal networks that represent a set of metabolic tasks for the study of specific cancer phenotypes. Then, the method MiNEA is used to perform minimal network enrichment analysis for the metabolic tasks.

3.2.2 Building cancer-specific models

The ever-increasing size of the human GEMs, including thousands of reactions and metabolites, hinders the analysis of the metabolic subnetworks of interest based on the changes observed in the expression data. Using a reduced model rather than the corresponding GEM facilitates the study of specific parts of metabolism and the comparison of metabolic alterations in specific pathways across physiologies. In addition, the computational complexity dramatically decreases when using a reduced model, enabling access to new results.

We used the redHUMAN workflow established in Chapter 2, to generate a thermodynamically curated reduced-version of the human GEM Recon 3D [33] around 11 metabolic subsystems that have been reported to be altered in cancer cells, namely,

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 (Materials and Methods). The systematic framework used to reconstruct the redHuman reduced model ensures its consistency with Recon 3D, the corresponding complete GEM. The redHuman reduced model is used as a generic metabolic model to build context-specific models for three cancer types, namely, breast cancer, colon cancer, and ovarian cancer, by integrating *omics* data from the NCI60 cancer cell lines following the previously defined workflow.

The human GEM Recon 3D was built as a model not specific for any cell-type nor tissue; thus, it includes a generic collection of biochemical reactions encoded in the human genome. This non-specificity is inherited by the redHuman reduced model. However, each tissue-specific cell type only expresses a portion of those genes, which encode only the enzymes that characterize the metabolism of the cell type. Therefore, leaving in the model all the alternative enzymes that are expressed in the human genome gives too much flexibility to the network miss-representing the specific physiology under study. Furthermore, on each tissue, cells have access to specific nutrients, and they secrete certain compounds to the extracellular medium. We captured these characteristics in the context-specific metabolic models by (i) defining the network topology according to the *omics* data and the genes that are expressed in the specific tissue, and (ii) defining the network physiology by constraining the metabolic reaction rates according to the expression data.

Within this approach, we identified deregulation in three levels: (i) gene deregulation, as it is defined by the transcriptomics data, (ii) enzyme deregulation, as the hypothetic reactions that are deregulated based on the transcriptomics data and the gene-protein-reaction (GPR) rules from the metabolic model, and (iii) reaction deregulation, as the deregulated reactions inferred from the workflow. In the following discussion, we will use these three deregulation terms distinctively as gene deregulation, enzyme deregulation, and reaction deregulation.

i. Defining the network topology for the specific cancer type

An essential aspect that defines the metabolism inside the cells is the ability of cells to transport metabolites from the extracellular and across intracellular compartments.

Given the genetic profile of each tissue, a specific set of transporters are expressed. A well-studied case is the transport of glucose [34], for which there exist two main families of transporters: the sodium-glucose linked transporters (SGLTs) and the facilitative glucose transporters (GLUTs). Among them, GLUT1 and GLUT2 are mainly expressed in hepatocytes, GLUT4 is expressed in heart, skeletal muscle, adipose tissue, and brain, and SGLT1 is expressed in intestinal cells. Similarly to glucose, other metabolites, including lactate, amino acids, and inorganics, can be transported by several enzymes. The tissue-specificity of these transporters is not yet well characterized at a genome-scale, and thus we derived a computational approach to describe the transport of metabolites across compartments in the context-specific models.

First, we integrated *omics* data from the breast, colon, and ovarian NCI60 cancer cell lines into each corresponding cancer-specific model. The data included concentrations of 115 medium metabolites, the uptake and secretion rates for those extracellular metabolites, transcriptomics data for 21212 genes, and doubling times for each cancer cell line (Materials and Methods). We integrated exo-metabolomics for 91 metabolites and exo-fluxomics values for 83 reactions into the models using the TFA formulation (Materials and Methods). We additionally constrained the growth rate based on the doubling time of each cancer type, the ATP maintenance reaction rate to 1.07mmol/(gDW·h) [35], and the oxygen uptake to 2 mmol/(gDW·h) [12]. The *omics* data allowed to constrain the space of reaction rates and metabolite concentrations in agreement with the corresponding physiology. By integrating the physiological conditions described by the metabolomics and fluxomics data into the generic reduced model, we generated three cancer models for breast, colon, and ovarian cancers.

Then, we defined in the cancer-specific models the transport reactions specific for each cancer type based on the expression profile and the metabolic requirements of the models. We first ensured that the electron transport chain reactions (classified as transporters as they pump protons between the mitochondria and the inner-mitochondrial membrane), as well as the transport reactions for small metabolites, remained in the models (Materials and Methods).

Next, we accounted for the cancer-type-specificity of the transport reactions by identifying the enzymes associated with expressed genes (Materials and Methods), assuming that if the gene is expressed, then the enzyme may be available, and the

corresponding reaction may take place. The generic redHuman reduced model has a total of 1377 transport reactions. Based on the expression of the genes in the transcriptomics data of the NCI60 cell lines and the gene-protein-reaction rules in the metabolic model, the enzymes catalyzing 1026 transport reactions are expressed in the three cancers but with a different deregulation profile on each cancer type (Figure 3.2 A and B). Colon cancer has the highest number of reactions associated with upregulated enzymes (256 reactions), followed by ovarian cancer (222) and breast cancer (211). On the contrary, breast cancer contains almost twice the number of transport reactions associated with downregulated enzymes than the other two cancers (183 vs. 87 in colon cancer and 102 in ovarian cancer). The number of transport reactions associated with not-deregulated enzymes is relatively similar across cancer types (between 683 and 702 reactions).

Finally, we used the redHuman reduced model to identify the additional transport reactions for which we could not assign expression data, and that should be active in order to allow the metabolic model to meet the observed growth phenotype. To this end, we formulated a mixed-integer linear programming (MILP) optimization problem that identified, from the 351 transport reactions that lacked gene-protein-reaction associations in the models, the minimum number of transport reactions required by each cancer model for the synthesis of biomass and we enumerated all the alternative sets of minimum size (Materials and Methods).

We identified 21 transport reactions that should be included in all the cancer models. These reactions transport mainly intermediates for nucleotide synthesis and cholesterol synthesis between cytosol and endoplasmic reticulum, deoxynucleotides from the cytosol to the nucleus, and intermediates of amino acid metabolism and citric acid cycle between extracellular and cytosol (Table S3.1). Moreover, the breast cancer model requires the transport of four additional compounds, namely, isocitrate, glucuronate, and creatinine between extracellular and cytosol, and L-isoleucine between mitochondria and cytosol. The colon cancer model needs transport reactions for glucuronate, creatinine, glyceraldehyde, adenosyl-homocysteine, and xanthine between extracellular and cytosol. Finally, the transport of L-isoleucine between mitochondria and cytosol, and adenosyl-homocysteine, xanthine, and dimethylglycine between extracellular and cytosol were required in the ovarian cancer model (Table S3.1).

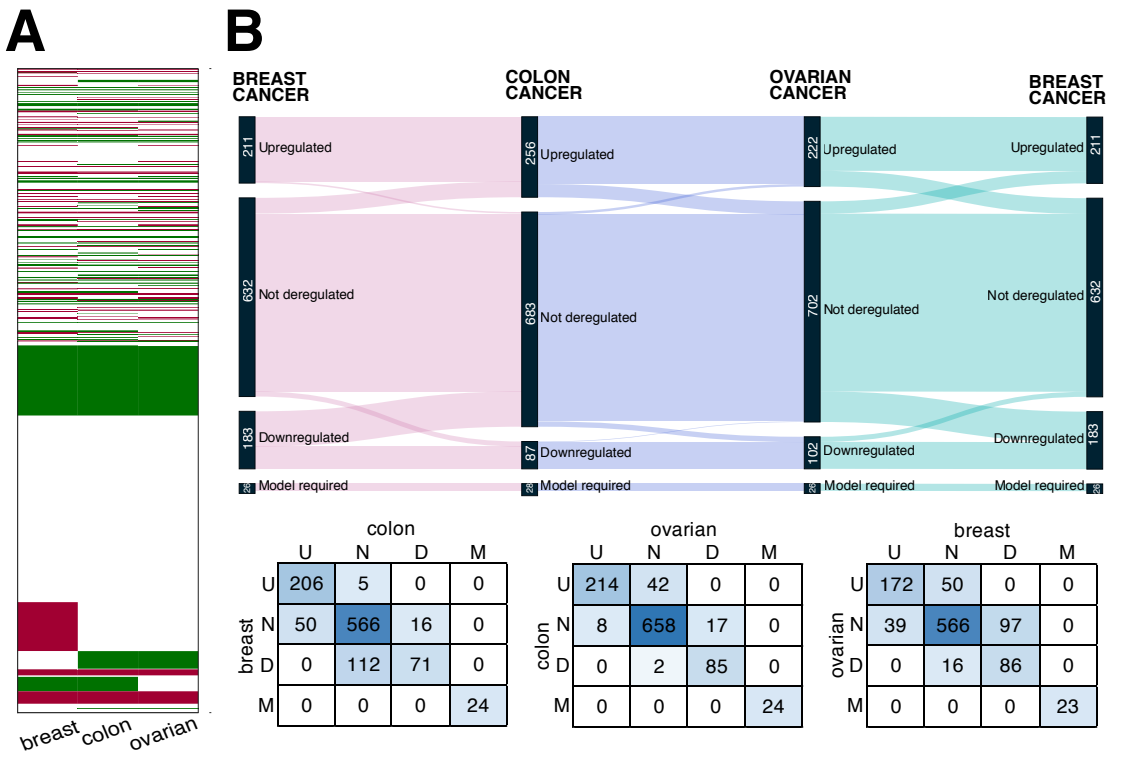


Figure 3.2. **Analysis of cancer-specific transport reactions.** (A) Expression profile for the enzymes of 1026 transport reactions across the three cancer types, breast cancer, colon cancer, and ovarian cancer. Green represents upregulated enzymes, red downregulated enzymes, and white those that are not deregulated. (B) Transport reactions kept in the cancer models based on gene expression and the model requirements for growth. Overlap of transport reactions per cancer type based on their deregulation and the model requirements (U: upregulated, N: not deregulated, D: downregulated, M: required by the model).

Following this pipeline, we constrained the topology of the cancer-specific models regarding transport reactions to 1091, 1092, and 1091 transport reactions for breast cancer, colon cancer, and ovarian cancer, respectively.

Overall, this analysis highlights the differences in transport reactions across cell types, and more importantly, it constrains the network topology of the corresponding model to the specific cell type, removing alternative pathways that are not active in the specific physiological conditions.

ii. Defining the cancer-specific physiology in the model

The physiology of each cell type is defined by a specific gene expression profile, which can be identified with transcriptomics data. The set of genes that are expressed on each cell type encode for particular enzymes that catalyze specific metabolic pathways.

Furthermore, quantitative transcriptomics data give information on the level of expression of the genes. We mapped this data to investigate the deregulation of the metabolic enzymes, assuming that a higher expression in the genes would lead to a higher quantity of metabolic enzymes and, therefore, to a higher flux through the metabolic pathways that these enzymes catalyze. Note here that if proteomics data are available, the same procedure can be followed using proteomics instead of transcriptomics data.

We used a recently developed method named TEX-FBA [36] to constrain the metabolic pathways in the models according to the transcriptomics profile. Given the transcriptomics data from the NCI60 cell lines for each cancer type, we derived the corresponding gene expression profiles per cancer-type (Materials and Methods). Then, using the gene-protein-reaction (GPRs) rules from the model, we assigned the consequent deregulation to the enzymes, and we maximized the number of reaction rates that could be constrained simultaneously to concur with the enzyme deregulation profile (Materials and Methods). The flexibility of the metabolic models allows us to enumerated alternative solutions that maximize the consistency between the deregulation of measured genes and simulated-reaction deregulation. We then identified the reactions that consistently appear in every alternative solution. We assigned with higher confidence that a reaction will be deregulated if it appears in all, or in a large number, of alternative sets.

Based on the transcriptomics data for breast cancer, 377 reactions were catalyzed by over-expressed enzymes and 186 reactions by lowly-expressed enzymes, resulting in a total of 563 reactions associated with deregulated enzymes. The stoichiometry, together with the metabolic demands of the pathways, introduce additional constraints to the reaction rates, often impeding to accommodate all the enzyme deregulations at the same time in the network. In the breast-cancer-specific model, a maximum of 559 out of the 563 reactions associated with deregulated enzymes could be constrained simultaneously to agree with the corresponding deregulation, and there exist 16 alternative sets of 559 reactions associated to deregulated enzymes. Finally, the rates of the 555 reactions that were part of all the alternatives are constrained in the breast-cancer-specific model to map the deregulation of their enzymes; specifically, 374 upregulated and 181 downregulated. We then clustered these 555 reactions based on their associated subsystems (Figure 3.3). Among the most upregulated subsystems, we

found the amino sugar metabolism, the nucleotide related pathways, the citric acid cycle, the oxidative phosphorylation pathway, and glycolysis. On the contrary, histidine metabolism was the most downregulated pathway which correlates with recent studies that report that a diet rich in histidine, and an increased histidine catabolism, increase the efficacy of the anticancer drug methotrexate [37, 38]. Moreover, we observed subsystems that contain both upregulated and downregulated reactions, as it is the case for tyrosine metabolism, tryptophan metabolism, urea cycle, transports, and pyruvate metabolism, among others.

Looking at the transcriptomics data, it would be hard to assign deregulation to the oxidative phosphorylation subsystem, as 33% of the reactions are associated with upregulated enzymes and 22% to downregulated enzymes (Figure 3.3 A). However, the integrative analysis provides strong evidence that the oxidative phosphorylation subsystem should be upregulated and interestingly 25% of enzymes for which we could not assign deregulation based on the transcriptomics, the flux balance solution suggests that they should be upregulated. Similarly, the transcriptomics data only suggest the deregulation of 44.7% of the enzymes in the glycolysis pathway. However, our analysis proves that 63.6% of the reactions in glycolysis are upregulated, correctly predicting the upregulation of the metabolic glycolytic fluxes characteristic of the Warburg effect.

The same procedure was used to define the reaction rates for the context-specific models for colon and ovarian cancer. In the case of colon cancer, the rates of 560 out of 568 reactions associated with de-regulated enzymes (464 up- and 104 down-regulated) could be simultaneously constrained according to their deregulation. From those 560, 559 reactions were consistent across 48 alternatives. The transcriptomics data for ovarian cancer associated 338 reactions from the ovarian cancer model to upregulated enzymes and 95 reactions to downregulated enzymes. In this case, a maximum of 431 reactions could be constrained to agree with the deregulation of their enzymes. After analyzing the consistency across the 4 existing alternatives, 429 reaction rates were constrained in the model to operate according to the deregulation of their enzymes.

Analyzing how the models can accommodate the fluxes to agree with the expression data, we observed that the three cancer types mostly upregulated the central carbon pathways, glycolysis, citric acid cycle, and pentose phosphate pathway, as well as, the nucleotide metabolism and alanine, aspartate and arginine and proline metabolism

(Figure 3.3) which are well-known pathways that are upregulated in cancer cells [39-42]. Interestingly, colon cancer presents a higher upregulation of folate metabolism than breast and ovarian cancer. It has been previously reported that folate metabolism plays an important role in colon cancer [43], and has been studied as a possible target for colon cancer [44, 45]. Regarding the downregulated pathways, all three cancers mostly downregulate genes associated with histidine metabolism and lysine metabolism [46]. Colon cancer shows a higher downregulation of the metabolism of aromatic amino acids (phenylalanine, tyrosine, and tryptophan), which is in agreement with previous studies [47] and an upregulation of the starch and sucrose metabolism which is not deregulated in the other two cancers. Dietary sugars, such as sucrose, were associated with an increased risk of colon cancer [48].

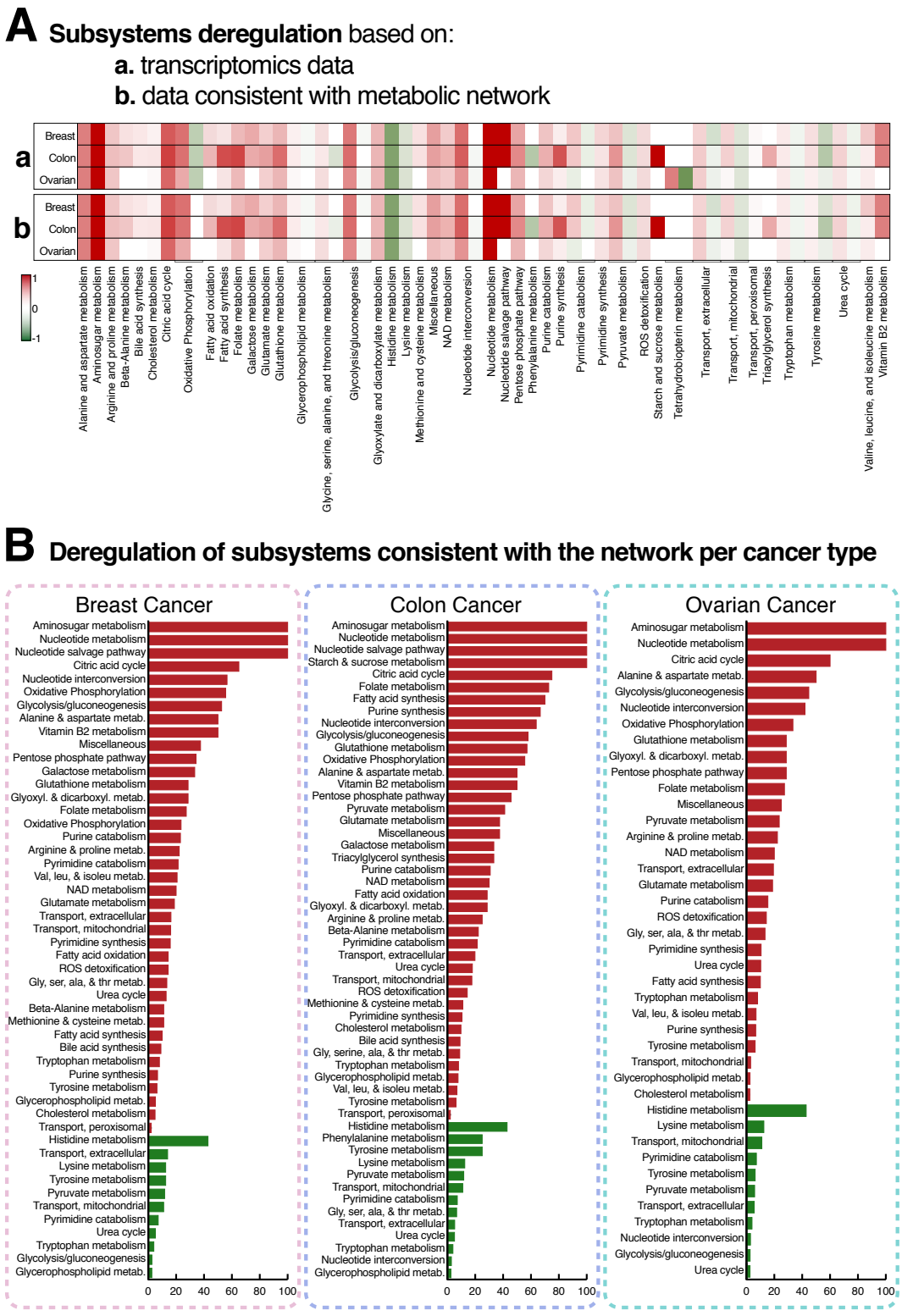


Figure 3.3. **Deregulation of subsystems in the context-specific models**. (A) Comparison of deregulated subsystems considering all the transcriptomics data and considering only the deregulated reactions that can have a metabolic rate consistent with the transcriptomics data and are common in all alternatives. (B) Percentage of deregulated subsystems based on the common

reactions across alternatives whose reaction rates can be consistent with the transcriptomics data for each cancer type.

This workflow allowed us to integrate expression data in the metabolic models not only qualitatively, that is, knowing which are the enzymes present, but also in a quantitative form by constraining the fluxes of the network according to the deregulation of the genes that code for the enzymes that catalyze them. Furthermore, we have developed a systematic procedure to analyze the consistency of the data with the metabolic flow and explain with the network the metabolic reasons for the deregulation of the genes in the corresponding cancers.

Moreover, based on the transcriptomics data, we usually focus on genes that are up- or down-regulated, and we assume with a certain confidence that the corresponding enzymes and metabolic pathways will be deregulated accordingly. When some genes associated with a pathway are upregulated, and some other genes are downregulated, we cannot immediately assign a specific deregulation to the pathway. The integration of the data in the GEMs and the workflow here proposed allows to identify fluxes that must be deregulated in order to preserve the metabolite mass balances when only a subset of them is considered to be deregulated based on metabolomics, proteomics and transcriptomics data.

3.2.3 Cancer phenotype analysis with metabolic models

As a result of applying the developed workflow, we generated three models representing the metabolism of breast, colon, and ovarian cancers, respectively (Figure 3.4 A). We used these cancer models to study the metabolic similarities and differences across the corresponding cancer physiologies.

We first performed thermodynamic-flux variability analysis while accounting for the expression constraints, on each cancer-specific model to characterize the space of reaction fluxes (Figure 3.4 B and Figure S3.1). Interestingly, the three cancer models present clear differences in their reaction rate variability. In general, colon cancer has a higher reaction rate flexibility, which corresponds to a higher growth rate. Reactions such as PFK or DPGM do not exhibit a big difference in flux across the three cancer types. However, some reactions, including PGK, FUMm, PCm, GTHOm, and ORNTArm

present evident differences across their fluxes. There are cases as GAPD where the fluxes across the three cancers do not overlap.

Moreover, there are reactions whose directionality differs among the three cancer types, for example, the reaction r0381, which converts hypotaurine to taurine and reduces NAD^+ to NADH is unidirectional in breast and colon cancer converting taurine to hypotaurine, and it is bidirectional in ovarian where hypotaurine can also be converted to taurine. Taurine has shown tumor attenuating effects in breast cancer and has been reported as a crucial metabolic pathway for breast cancer [49, 50]. Furthermore, it has been proven that it conducts apoptosis in colon cancer cells.

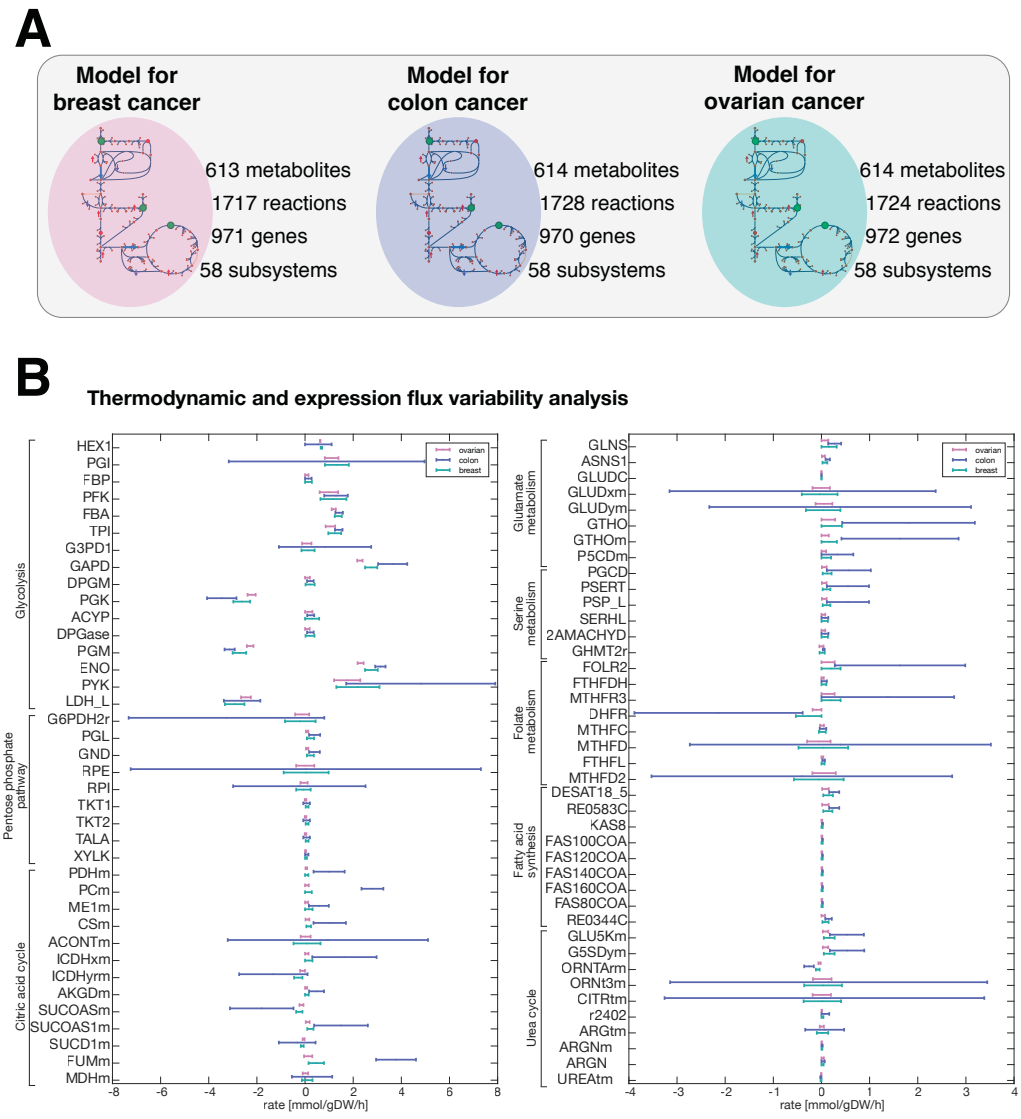


Figure 3.4. **Cancer phenotype analysis**. (A) Cancer-specific models. Number of metabolites, reactions, genes, and subsystems in each cancer-specific model. (B) Reaction rate variability analysis for the three cancer types.

We next analyzed the essential components of the models by performing gene and enzyme essentiality analysis on the three cancer models (Materials and Methods). We identified 32 genes that are commonly essential across cancer types (Figure 3.5 A). These genes code for enzymes mainly from the subsystems nucleotide interconversion and cholesterol metabolism. In addition, we identified GMPS as an essential gene in the breast cancer and colon cancer models but not in the ovarian cancer model. This gene is involved in the *de novo* synthesis of guanine nucleotides, which have been reported to be essential for DNA and RNA synthesis, and they also provide GTP, which is involved in a number of cellular processes important for cell division [51]. Furthermore, ALDH18A1 was found to be essential only in the colon cancer model. ALDH18A1 is a bifunctional enzyme that converts glutamate to glutamate 5-semialdehyde, an intermediate in the biosynthesis of proline, ornithine, and arginine [51]. ALDH18A1 has been reported as a specific gene for colon and intestinal tissue [52].

Single gene essentiality identifies the genes that are necessary to simulate growth in the model. However, some enzymes are coded by several genes, and the presence of one of them is enough to transcribe the enzyme. In that case, the genes will not be essential, but the enzyme may be necessary for a specific physiology when a pathway must be functional. We characterized these indispensable parts of the network by performing enzyme essentiality, that is, analyzing the enzymes that should be active in the corresponding cancer physiology to sustain growth.

A total of 218 enzymes showed to be essential in the network to display the specific phenotypes (Figure 3.5 B). Specifically, 162 enzymes were essential in the breast cancer physiology, 201 enzymes were required in colon cancer, and 130 enzymes in the case of ovarian cancer. Additionally, we identified eight enzymes that are essential in breast cancer but not in the other two cancer physiologies. These enzymes catalyze reactions from the pentose phosphate pathway, from serine metabolism, from methionine metabolism, and two extracellular transport reactions, suggesting that these pathways are more important for breast cancer than in the case of the other two cancers. In the case of colon cancer, 49 out of the total 201 (24%) of the essential enzymes were specific to this cancer. These specific essential enzymes include mainly extracellular transports, and reactions from nucleotide interconversion, tyrosine metabolism, tryptophan metabolism, and lipid-related subsystems, among others. Finally, we identified two enzymes that were specifically essential for ovarian cancer. One of these two enzymes

catalyzes a reaction from the serine pathway, and another one catalyzes an extracellular transport reaction.

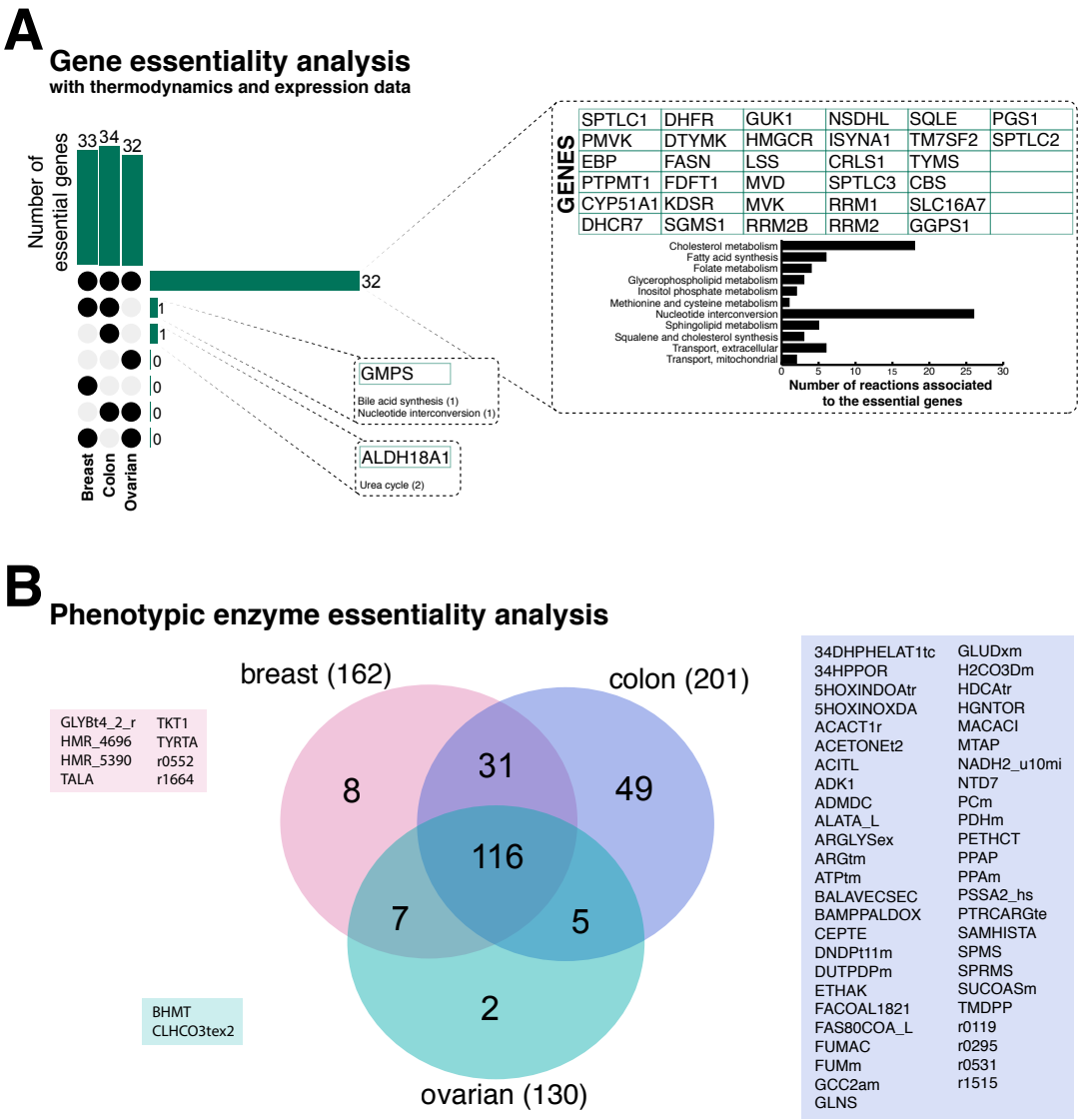


Figure 3.5. **Essentiality analysis**. (A) Gene essentiality for the three cancer types performed, taking into account omics data, thermodynamics, and expression constraints. (B) Enzyme essentiality analysis for the three cancer types with the data imposed in the network for thermodynamics, metabolite concentrations, reaction rates, and consistency with transcriptomics data.

3.2.4 From pathways to minimal networks

The genetic and regulatory changes that accompany disease development and progression reflect in modifications in the activity of the enzymes and consequently, in altered metabolic profiles that impact the pathways and the overall functioning of the cell. The metabolic functions of cells can be represented with metabolic tasks such as the synthesis of a metabolite or the activity of specific reactions. An example of a metabolic task is the synthesis of phosphatidylserine, which is required for the formation of cellular membranes. The classical pathway for the synthesis of phosphatidylserine includes four steps starting from glycerol 3-phosphate and acetyl-CoA (Figure 3.6 A). The deregulation of any of these four steps compromises the synthesis of the phospholipid. However, the deregulation of the pathways providing the precursors can also impact its synthesis.

We considered all the upstream deregulation effects by identifying the minimal set of reactions required to be active for the synthesis of phosphatidylserine. This minimal network comprehends not only the classical four steps but also the reactions required for the synthesis of the fatty acids that form the tails of the phospholipid and the synthesis of the main precursors, glycerol 3-phosphate and acetyl-coA from glycolysis and the citric acid cycle, respectively (Figure 3.6 A). Indeed, the deregulation of these additional reactions and the pathways that provide the precursors affects the overall synthesis of the phospholipid, and we would not have assigned this deregulation to the metabolic task if we had looked only at the classical pathway.

Aiming to study the deregulation of metabolic tasks for the different types of cancer, we used MiNEA [53], a method to generate minimal networks (MiNs) required for metabolic tasks and to enrich the MiNs based on the deregulation of the genes associated to the reactions that compose them. Following the MiNEA workflow, we defined a set of metabolic tasks that describe seven phenotypes in cancer cells (Materials and Methods), including the Warburg effect (production of lactate), glutamine addiction (production of glutamate), reprogramming of energy metabolism (production of ATP through the electron transport chain), stress response (production of the reactive oxygen species superoxide anion and hydrogen peroxide), reprogramming of pentose phosphate pathway (production of ribose 5-phosphate), altered serine pathway (catabolism of serine and production of glycine) and phospholipid synthesis (production of phosphatidyl-serine, phosphatidyl-choline, and phosphatidyl-myo-inositol).

Then, we used the cancer-specific models to generate thermodynamically feasible minimal networks (MiNs) associated with each metabolic task, and we enumerated all the corresponding alternative reaction sets of minimal size that capture the ability of cells to use alternative routes to synthesize the target metabolites (Materials and Methods). For each metabolic task, we identified the common reactions across alternatives, known as high-frequency reactions (HFRs), which represent the constitutive set of reactions that must be active to perform the task within a minimal network (Table 3.1).

For all the cancer types, glutamine, serine and phosphatidyl-choline presented the longest MiN and ribose-5-phosphate the shortest MiN. The metabolic tasks with the highest number of alternatives were ribose-5-phosphate in breast cancer, glycine in colon cancer and ATP in the case of ovarian cancer, implying that the cells present more flexibility to perform these tasks. Analyzing the number of HFRs per metabolic task, we observed that more than 40% of the reactions are conserved across alternatives for all the metabolic tasks. Colon cancer has the highest percentage of common reactions across MiNs, but also the least number of alternatives. Based on these results, we could hypothesize that colon cancer cells have less flexibility to perform these metabolic tasks. On the other hand, ribose-5-phosphate in breast cancer, lactate in colon cancer and superoxide anion in ovarian cancer have the lowest percentage of common reactions across alternatives, suggesting a higher diversity for these tasks in comparison with the other metabolic tasks here analyzed.

As an example of a minimal network, we show the subnetworks that represent the Warburg effect phenotype (lactate production) for the three cancer types (Figure 3.6 B). As expected, the MiN that represents the Warburg effect contains active reactions mainly from glycolysis. Interestingly, the results reveal that to have an active glycolysis pathway, additional reactions are required, including reactions from the pentose phosphate pathway and from the citric acid cycle. It is important to note here that the deregulation of these additional reactions will affect the production of lactate.

Table 3.1. **MiNs per cancer-type for different phenotypes**. MiN: Minimal network, HFRs: High-frequency reactions. Phenotypes and metabolic tasks associated: Warburg effect (lactate), glutamine addiction (glutamate), stress response (superoxide anion and hydrogen peroxide), energy metabolism (ATP through the electron transport chain), serine pathway (serine and glycine), reprogramming of the pentose phosphate pathway (ribose-5P) and phospholipid synthesis (phosphatidyl-serine, phosphatidyl-choline, and phosphatidyl-inositol).

Cancer type	Phenotype	Metabolic Task	MiN		HFRs	
			Size	Number of alternatives	Number	Percentage
Breast	Warburg effect	L-lac	71	69	53	74.65
	Glutamine addiction	L-glu	87	20	64	73.56
	Stress response	O_2^-	55	96	40	72.72
		H_2O_2	38	89	21	55.26
	Energy Metabolism	ATP & ATPS	63	221	46	73.0
	Serine pathway	L-ser	87	206	59	67.8
		L-gly	46	81	37	80.4
	PPP	rib-5p	31	288	14	45.2
	Lipid Metabolism	PS	64	250	45	70.3
		PC	87	16	65	74.7
		PI	47	160	34	72.3
Colon	Warburg effect	L-lac	56	8	31	55.4
	Glutamine addiction	L-glu	93	1	93	100.0
	Stress response	O_2^-	60	8	50	83.3
		H_2O_2	43	32	38	88.4
	Energy Metabolism	ATP & ATPS	77	1	77	100.0
	Serine pathway	L-ser	94	7	64	68.1
		L-gly	33	154	23	69.7
	PPP	rib-5p	30	3	27	90.0
	Lipid Metabolism	PS	68	3	52	76.5
		PC	122	1	122	100.0
		PI	56	1	56	100.0
Ovarian	Warburg effect	L-lac	45	226	36	80.0
	Glutamine addiction	L-glu	91	110	69	75.8
	Stress response	O_2^-	50	202	26	52.0
		H_2O_2	58	240	49	84.5
	Energy Metabolism	ATP & ATPS	57	295	47	82.5
	Serine pathway	L-ser	101	1	101	100.0
		L-gly	52	21	37	71.2
	PPP	rib-5p	27	16	21	77.8
	Lipid Metabolism	PS	71	116	40	56.3
		PC	86	100	51	59.3
		PI	47	34	34	72.3

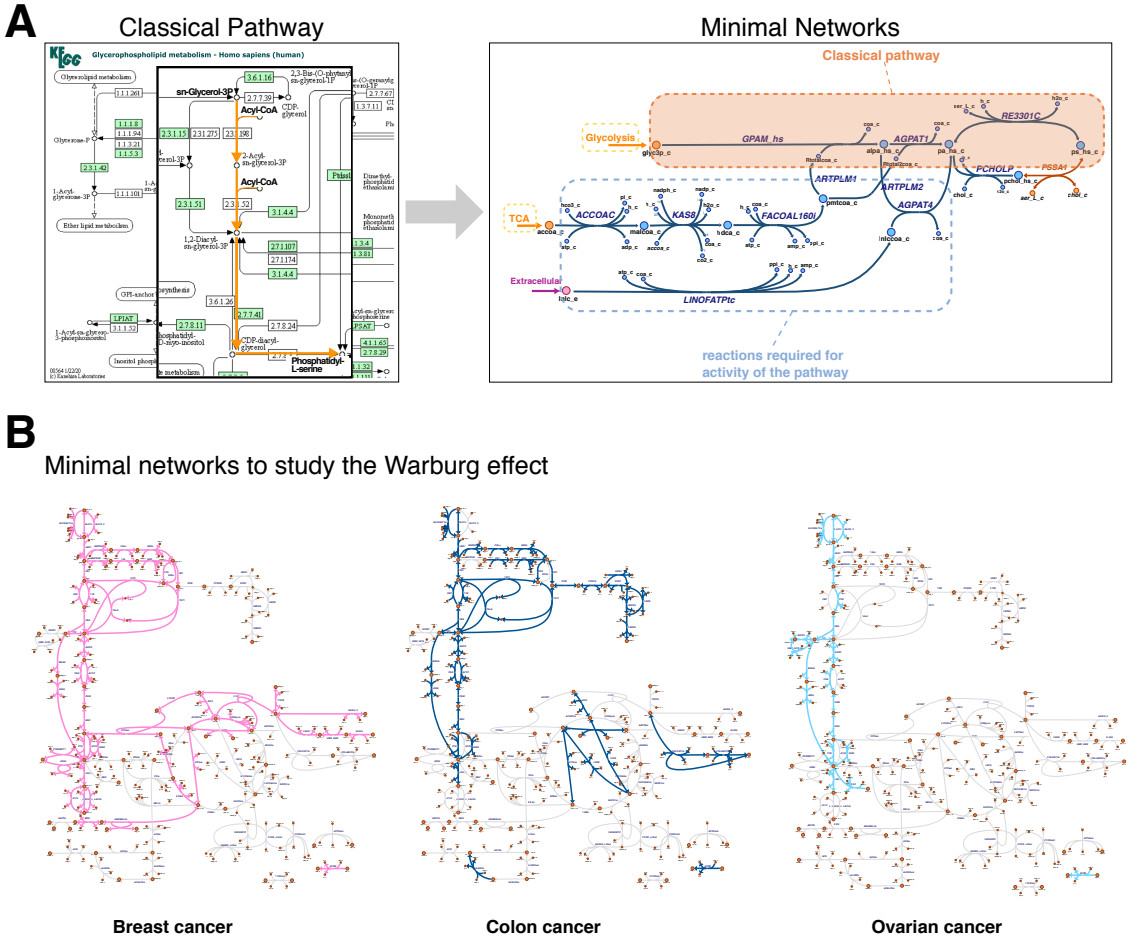


Figure 3.6. **Minimal Networks**. Minimal networks vs classical pathway. (A) representation of the classical pathway in KEGG for the synthesis of phosphatidyl-serine and the corresponding minimal network, which also includes the upstream pathways, in this case, glycolysis and TCA. (B) Minimal networks representing the phenotypes of the Warburg effect for breast, colon, and ovarian cancers. For the sake of simplicity in the visualization the networks do not include all the reactions from the MiN, but the central pathways that show the main differences among them.

3.2.5 Minimal network deregulation and enrichment analysis

A common approach to assign functionality to a set of genes from a pathway is to perform gene or pathway enrichment analysis, where statistical tests are used to determine the importance of the genes in the specific functional set. In this work, we performed instead minimal network enrichment analysis, using MiNEA to enrich the set of genes that belong to each minimal network (Materials and Methods) for each task. We assigned deregulation to the minimal networks by using the transcriptomics data and the GPR rules from the model. Then, the significance of the enrichment was tested using a hypergeometric test (Materials and Methods).

The results show that all the tasks, and the phenotypes, here analyzed are significantly upregulated for the three cancers, correlating with the reported evidence. Moreover, the production of ATP was the most upregulated task in breast and ovarian cancers while the production of L-lactate and phosphatidylserine were the most upregulated tasks in colon cancer (Figure 3.7). The analysis shows that the deregulation of the genes highly affects the reactions that are required for these metabolic tasks, allowing to assign with higher confidence deregulation to the metabolic tasks based on the deregulation of the corresponding genes.

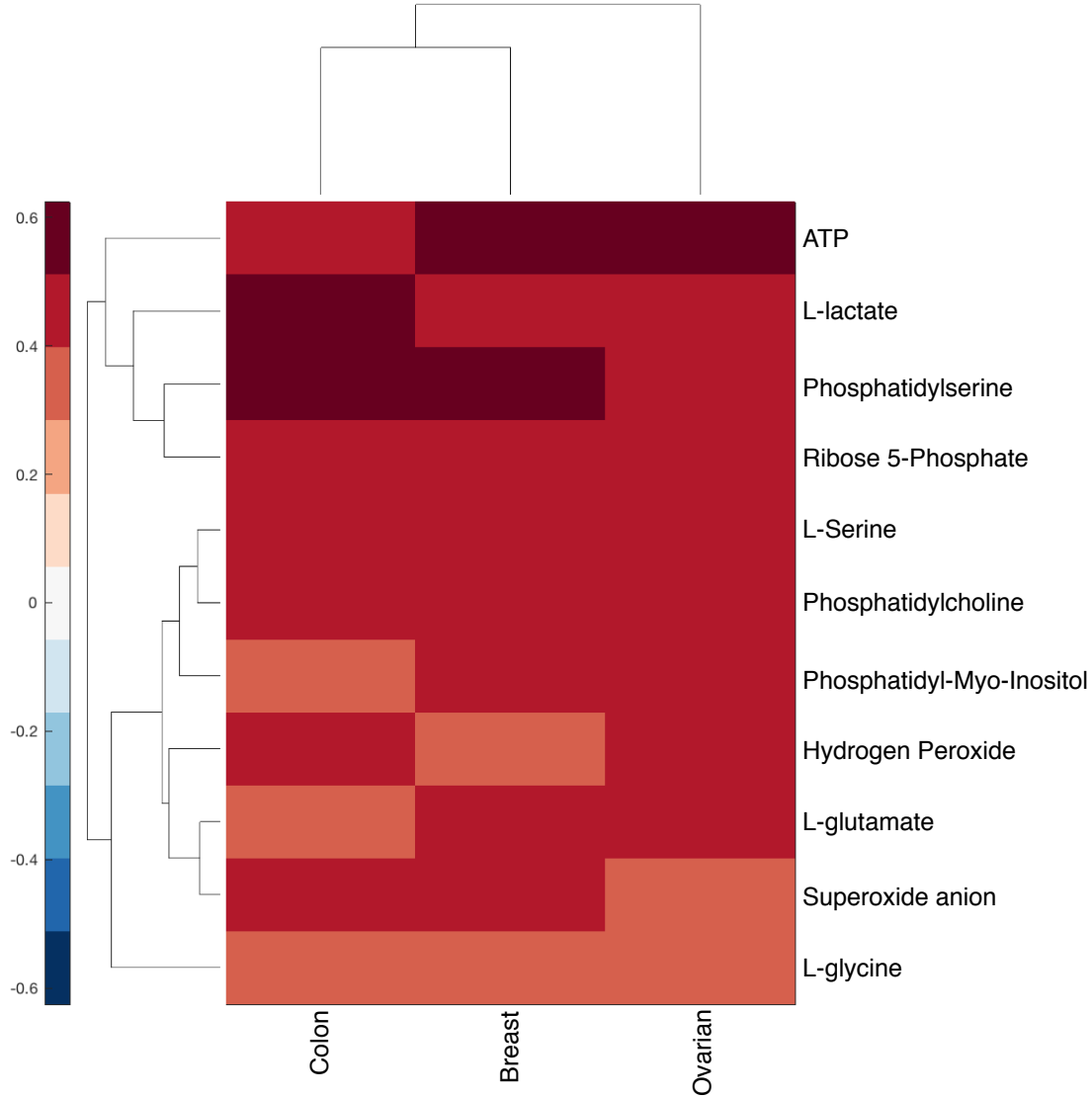


Figure 3.7. **Deregulation of metabolic tasks**. Most significantly deregulated (p<0.001) metabolic tasks for each cancer type.

3.3 Discussion

A context-specific model represents with high accuracy the intracellular metabolism of a specific type of cell under defined physiological conditions. Context-specific models are valuable tools for the study of the metabolic adaptation of cells to different extracellular conditions and to compare the intracellular metabolism of different types of cells, for example, cells from different tissues, or healthy vs. diseased cells. Different types of cells have different gene expression patterns, and consequently, different metabolic enzymes available, and thus they manifest differences in the metabolic pathways used to synthesize the metabolites required to perform cellular functions. To study these differences, we have developed a workflow to build context-specific metabolic models by integrating *omics* data and computational methods to identify the differences in the metabolic phenotype expressed by the different types of cells under study.

In this work, we generated cancer-specific models using as a scaffold a reduced version of a thermodynamically curated human genome-scale model, and we used the derived cancer-specific models to study the metabolism of cancer cells and the different metabolic pathways that the different cancer cells use to survive. The workflow to generate cancer-specific models includes a curation of the transport reactions based on the corresponding cancer expression data. We considered that all the reactions associated with expressed genes were part of the physiology. However, we further constrained the enzymes associated with lowly expressed genes to have lower activity and those associated with higher expressed genes to sustain higher activity. With this, we minimize the bias of removing reactions from the network, and we consistently constrain the metabolic flow to the corresponding expression data for the specific physiology.

The transcriptomics data is further used to translate the deregulation of the genes into deregulations in the metabolic reactions. The context-specific models here developed have proven to be powerful tools to infer deregulation in the metabolic pathways that could not be immediately seen in the transcriptomics data. Showing that by constraining a subset of reactions to match the measured transcriptomics data, the metabolic network imposes a deregulated flux in other pathways that allow the model to achieve the simulated phenotype.

The essentiality analysis performed with the context-specific models combined with experimental analysis can help to further develop the models. Furthermore, it can be used as a basis to generate hypotheses that can lead to the discovery of new targets for each cancer type and to understand why targeting a protein or gene for one cancer may not be valid for another cancer.

The generated cancer-specific models are used to analyze the deregulation of metabolic pathways in the different cancer types and to identify the functionality of the deregulations, in terms of metabolic tasks. Our approach allows to characterize the alternative pathways that cells may use to respond to changes in the environment. These alternatives are also used to overcome the lack of data, describing the possible scenarios that better describe the partial data.

The developed workflow has promising applications in the study of the different cellular phenotypes within the tumor microenvironment such as modeling cells from the surface of the tumor mass and the cells that are in the hypoxic areas to identify the differences in their metabolism triggered by the diverse accessibility to nutrients which compels cells to adapt their expression profiles to survive and function [54-57].

In this work, we have used cancer data from the NCI60 cancer cell lines based on the cancer type; however, an analogous analysis can be performed for different subtypes of a specific cancer type, for example, in the case of breast cancer, by generating models for the specific type of breast cancer. This will allow to analyze the metabolic signatures for the different breast cancer subtypes. Furthermore, the same workflow can be applied using other sets of cancer data containing metabolomics, fluxomics, proteomics, and transcriptomics, as well as using patient-specific data. In the absence of extracellular metabolomics and doubling time, as in the case of primary tissue data, an additional analysis would be performed to infer the nutrient requirements based on, for example, the expression of the genes and enzymes and the constraints in the metabolic model. Finally, the same pipeline can be used for any other tissue, condition, and disease.

Overall, the models here presented and the analysis performed can be used in combination with experimental studies to hypothesize the reasons for which cells change their expression profiles. Bringing the opportunity to explore common pathways across cell types, metabolic signatures, competition in the tumor microenvironment, and metabolic effects of the signal transduction. Moreover, it provides a step forward in the

search of clarifying the underlying reasons for cells to become carcinogenic and in the discovery of biomarkers and personalized cancer medicine.

3.4 Materials and Methods

3.4.1 Exometabolomics, transcriptomics data and gene expression per cell line

In this work, we use cancer *omics* data from the NCI60 cancer cell lines. We collected extracellular metabolomics and metabolite consumption and released values measured and reported by Jain et al. [32]. Concentrations and reaction rates were transformed into TFA units, that is, mol/L and mmol/(gDW·h), respectively. A conversion of 0.2 ngDW/cell was used [12]. *In-vitro* doubling times and normalized transcriptomics data were obtained from the NCI data repository (https://www.ncbi.nlm.nih.gov/sites/GDSbrowser?acc=GDS4296). Growth rates per cell line were calculated, dividing $\ln(2)$ over the corresponding doubling time.

Gene expression data were derived from the normalized transcriptomics data. For each gene, we analyze its population of transcriptomics across cell lines, and we classify the gene as under-expressed in the corresponding cell line, if its sample value is lower than the first quartile of the population, or over-expressed, if its sample value is higher than the third quartile of the population.

3.4.2 Cancer-type specific data

We define the cancer-type specific data considering the NCI60 cell lines corresponding to breast cancer (MCF7, MDA-MB-231/ATCC, MDA-MB-468, HS 578T, BT-549, T-47D), colon cancer (COLO 205, HCC-2998, HCT-116, HCT-15, HT29, KM12, SW-620), and ovarian cancer (IGR-OV1, OVCAR-3, OVCAR-4, OVCAR-5, OVCAR-8, NCI/ADR-RES, SK-OV-3). See Table S3.4 for all the cancer cell lines.

For each cancer type, we define the reaction rates and concentrations values as ranging from the lowest value to the highest value of the cell lines that correspond to the cancer type.

We defined the gene expression data per cancer type by analyzing the gene expression data across the corresponding cell lines. For each gene, if it is consistently deregulated

(up or down) in at least 80% of the samples, we assign the resulting deregulation to the gene in the specific cancer type.

3.4.3 Human generic metabolic model

The metabolic model is a reduction of the thermodynamically curated human genome-scale model Recon 3D [33], generated following a recently developed pipeline named redHUMAN. The reduction model was generated around 11 subsystems that have been reported to be altered in cancer, namely, 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. The model consists of 947 metabolites, 2316 reactions, and 1063 genes associated with the reactions. It contains a total of 63 subsystems, and 106 extracellular metabolites were connected to the initial subsystems considering all the alternative pathways of minimal size (parameter Smin in redGEMX in the redHUMAN workflow). Moreover, the model contains all the reactions that are required to produce the biomass building blocks, including minimal size subnetworks to minimal size plus three (parameter Sminp3 in lumpGEM [30, 31]).

Seeking to have a complete description of the metabolism in the model, we removed the lumped reactions and added the reactions of their corresponding subnetworks in the model.

3.4.4 Integrating context-specific metabolomics and fluxomics data

Metabolomics and fluxomics data for the specific cancer type are integrated as constraints in the model. In exact, we use the cancer type *omics* data to define the lower and upper bounds of the corresponding metabolite concentrations and reaction rates in the TFA formulation of the model.

3.4.5 Defining context-specific transport reactions

Having integrated the metabolomics and fluxomics data that characterize the tissue. The selection of the transports is performed as in the following steps:

1. Identify the transport reactions associated with expressed genes. These reactions will remain on the model. Assuming that the enzymes associated with expressed genes may be active.

2. Find the reactions associated with the electron transport chain and the transport of small molecules such as oxygen, protons, sulfate, phosphate, water, and carbon dioxide and the following inorganic molecules: ammonium, potassium, sodium, chloride, diphosphate, hydrogen peroxide, superoxide, nitric oxide (Table S3.2).
3. Assuming a network with m metabolites and n reactions. We formulate the following mixed-integer linear programming (MILP) problem to find the minimum number of the remaining transports that need to be part of the model to simulate growth.

$$\max \sum_{k=1}^{R^t} z_k$$

subject to:

$$\boldsymbol{S} \cdot \boldsymbol{v} = \mathbf{0} \quad (1)$$

$$v_i^{LB} \leq v_i \leq v_i^{UB}, \forall i \in R^M \quad (2)$$

$$\Delta_r \mathrm{G}_i' = \sum_{j=1}^{m} n_{i,j} \Delta_f \mathrm{G}_j'^{o} + RT \ln\left(\prod_{j=1}^{m} x_j^{n_{i,j}}\right), \forall i \in R^M, \quad (3)$$

$$\Delta_r \mathrm{G}_i' - M + M \cdot b_i^F \leq 0 \quad (4)$$

$$-\Delta_r \mathrm{G}_i' - M + M \cdot b_i^R \leq 0 \quad (5)$$

$$v_i^{F,R} - M \cdot b_i^{F,R} \leq 0 \quad (6)$$

$$b_i^F + b_i^R \leq 1 \quad (7)$$

$$b_k^F + b_k^R + z_k \leq 1, \forall k \in R^t \quad (8)$$

$$v_{biomass}^{LB} = c \cdot v_{biomas,max} \quad (9)$$

where $i = 1, \ldots, n,\ j = 1, \ldots, m$, $\boldsymbol{S}$ is the stoichiometric matrix, $\boldsymbol{v}$ are the net fluxes for all the reactions and v_i^F, v_i^R are the corresponding net-forward and net-reverse fluxes, so that, $v_i = v_i^F - v_i^R, \forall\, i = 1, \ldots, n.$ v_i^{LB} and v_i^{UB} are the lower and upper bounds, respectively, for the ith reaction. $\Delta_r \mathbf{G}'$ is the Gibb's free energy of the reactions defined in TFA [26, 27]. $n_{i,j}$ is the stoichiometric coefficient of compound j in reaction i; $\Delta_f \mathrm{G}_j'^{o}$ is the standard Gibbs free energy

of formation of compound j; x_j is the concentration of the compound j; R is the ideal gas constant, $R = 8.31 \cdot 10^{-3} \frac{KJ}{K\,mol}$, and T is the temperature. In this case, $T = 298\,K$. b_i^F and b_i^R are the binary variables for the forward and reverse fluxes of all the reactions (coupled to TFA). M is a big constant (bigger than all upper bounds). R^M denotes the set of reactions in the model and R^t denotes the set of transport reactions that are not expressed neither from the ETC nor transports for small molecules. z_k, b_k^F and b_k^R are binary variables to control the activity and the flux through the kth transport reaction of R^t. c is a constant to select the percentage of $v_{biomas,max}$. In this work, $c = 1$ as we want the model to have the capability to produce 100% of biomass. If $z_k = 1$, then the corresponding transport reaction cannot carry flux.

Equations (1) and (2) represent the constraints of the FBA problem, equations (3) to (7) represent the additional constraints required to form the TFA problem as defined in [26, 27], and equations (8) and (9) are the constraints to formulate the MILP that finds the minimum number of transport reactions required for growth.

4. Generate and characterize the alternative sets of minimum number of transports for the previous MILP. Identify from the alternative sets the transports that will remain in the context-specific model.
5. Build a model with the selected set of transports and the boundary and intracellular reactions from the generic model that can carry flux.

3.4.6 Thermodynamic-flux variability analysis

In this work, thermodynamic-flux variability analysis is performed in combination with the common assumption of minimization of resources, applied as minimization of the sum of fluxes which constrain the space of solutions to a more biologically relevant space [58, 59].

Assuming a network with n reactions, thermodynamic flux variability analysis is performed, for each reaction r_k ($k = 1,..n$) as it follows:

$$\min/\max \quad v_k$$

subject to:

TFA formulation: Eq. (1 – 7)

$$\sum_{i=1}^{n} v_i = V_{min}$$

$$v_{biomass}^{LB} = c \cdot v_{biomas,max}$$

where $v_{biomas,max}$ is the maximum growth for the model, $c = 0.9$, and V_{min} is the solution of the following MILP:

$$\min \sum_{i=1}^{n} v_i$$

subject to:

TFA formulation: Eq. (1 – 7)

$$v_{biomass}^{LB} = c \cdot v_{biomas,max}$$

where $v_{biomas,max}$ is the maximum growth for the model and $c = 0.9$.

3.4.7 Integration of context-specific expression data

We use TEX-FBA [36] and the transcriptomics data to constrain the reaction rates in the model. TEX-FBA works under the assumption that enzymes associated to over express genes are more active and the corresponding reactions carry more flux, and those enzymes associated to down regulated genes are less active and the reactions carry less flux. This method leaves certain flexibility in the fluxes of the reactions to account for the no one-to-one relationship between gene expression and enzyme activity due to post-transcriptional and post-translational modifications.

We found the maximum set of reactions whose fluxes can simultaneously be constrained to agree with the expression data as it follows:

1. Define reaction deregulation based on the gene expression data (transcriptomics) as defined in [36].

2. Perform thermodynamic-flux variability analysis in the context-specific models which have metabolomics and fluxomics data and the context-specific transports.

3. Optimize the following MILP problem:

$$\max \sum_{k \in (R_H \cup R_L)} z_k$$

subject to:

TFA formulation: Eq. (1 - 7)

$$v_h^{LB} = \alpha \cdot (v_{h,max} - v_{h,min}) \cdot z_h + v_{h,min} \cdot (1 - z_h)\,, \qquad \forall h \in R_H \qquad (10)$$

$$v_h^{UB} = v_{h,max} \qquad \forall h \in R_H \qquad (11)$$

$$v_l^{LB} = v_{l,min} \qquad \forall l \in R_L \qquad (12)$$

$$v_l^{UB} = \beta \cdot (v_{l,max} - v_{l,min}) \cdot z_l + v_{l,max} \cdot (1 - z_l), \qquad \forall l \in R_L \qquad (13)$$

where, R_H and R_L denote the set of upregulated and downregulated reactions respectively. For each reaction, $(v_{k,min}, v_{k,max})$ is the range obtained by performing thermodynamic-flux variability analysis. α and β are parameters to define how much percentage of flux is demanded. In this work, we used $\alpha = 0.7$ and $\beta = 0.3$. z_k are binary variables that control the activity of the reactions. If $z_k = 1$ then the kth reaction rate is forced accordingly to the deregulation. On the contrary, if $z_k = 0$ then the reaction is not forced to be consistent with the corresponding deregulation.

As a result, TEX-FBA maximizes the number of reactions whose fluxes can be constraint to $(\alpha v_{k,max} - \alpha v_{k,\min}, v_{k,max})$ if the reaction is upregulated and to $(v_{k,min}, \beta v_{k,max} - \beta v_{k,min})$ if it is downregulated.

4. We then generate alternative sets of maximum consistency with the expression data.
5. Only those reactions that are present in all the alternatives are finally constraint to be consistent with the data.

3.4.8 Gene and enzyme essentiality analysis

For this analysis, we performed thermodynamic flux variability analysis (TVA) in the models with the expression constraints added with TEX. Then we used the values of the TVA to constrain the lower and upper bounds of the fluxes in the corresponding cancer models without the expression constraints. These models were used to perform essentiality analysis.

Gene essentiality analysis was performed as it follows. We first relaxed the bounds to not impose forced reaction rates in the models. That is, we relaxed the lower bounds to zero for reactions operating in the forward direction, and the upper bounds to zero for reactions operating in the reverse direction. This is done to avoid essentiality imposed by reaction rates instead of genes. Then, we performed an individual knock-out of each gene and identified the reactions affected by evaluating the gene-protein-reaction (GPR) rules. The flux through those reactions is blocked, and the model is tested for growth. If the model cannot simulate growth, then the corresponding gene is considered essential.

Enzyme essentiality analysis was based on the GPR rules. First, we identified the enzymes whose reactions have the same GPR rules. That is, we considered that two enzymes would be the same if they have the same gene rules. We then blocked the flux through the reactions catalyzed by each enzyme, and we analyzed if the model can simulate growth. In the case it cannot, we considered the enzyme essential.

3.4.9 Formulating metabolic tasks for cancer physiology

In order to sustain a rapid proliferation rate and survive in the tumor microenvironment, cancer cells undergo a series of transformations, including reprogramming of metabolism. These modifications manifest in a variety of metabolic phenotypes that have been exploited as targets for cancer therapies in the last years [2, 9, 60]. One of the most studied cancer phenotypes is the Warburg effect, first described by Otto Warburg in the 1920s when he observed that cancer cells increase their uptake of glucose and perform aerobic glycolysis [14, 61, 62]. Besides the high glucose intake, cancer presents addiction to the non-essential amino acid glutamine, which contributes as a nitrogen source for nucleotide synthesis and, as energy and anaplerotic source to replenish the citric acid cycle [63-65]. Additionally, cancer cells experience a reprogramming of energy metabolism to support mitochondrial activity [66-68]. The alterations in mitochondrial

functions increase oxidative stress and ROS levels, which promote tumor growth and progression [69]. Together with glycolysis, cancer cells upregulate the pentose phosphate pathway, which is a source of ribonucleotides and NADPH [70], and serine metabolism, which contributes to the one-carbon metabolism [71-74]. Moreover, tumor cells show a reprogramming in lipid metabolism increasing lipogenesis and affecting the composition of the cellular membranes [75-77].

We defined the metabolic tasks by assigning a set of metabolites that represent each cancer phenotype. Consequently, we defined the metabolic tasks of the Warburg effect as the production of lactate. Production of glutamate represents the metabolic task for the phenotype glutamine addiction. To study the reprogramming of energy metabolism, we defined the metabolic task as the production of ATP through the electron transport chain. Oxidative stress was studied through the production of superoxide anion and hydrogen peroxide. The reprogramming of the pentose phosphate pathway is analyzed through the production of ribose 5-phosphate. Altered serine metabolism is represented with the production of serine and glycine. Finally, the reprogramming on lipid metabolism was studied through the production of the phospholipids phosphatidyl-serine, phosphatidyl-choline, and phosphatidyl-myo-inositol.

3.4.10 Generating Minimal Networks for the metabolic tasks

For each metabolic task, we generate minimal networks, that is, the minimum set of reactions from the model that needs to be active to satisfy the specific metabolic task (synthesis of a metabolite or activity of a reaction). Note here that the models have been thermodynamically curated, and we use the TFA formulation guaranteeing that all the pathways are thermodynamically feasible.

In this work, the minimal networks are found using the following MILP problem:

i. Constrain the reaction rate bounds to the values obtained with the flux variability analysis from the model with TEX-FBA constraints.

ii. Relax the lower bounds to zero, to avoid having reactions forced. This step guarantees that only reactions required for the task appear as a solution of the MiN. If a reaction was forced to carry flux, it would be part of every MiN.

iii. Add a sink reaction that represents the production of the metabolite required for the metabolic task.

iv. Optimize the following MILP:

$$\max \sum_{i=1}^{R^M} z_k$$

subject to:

TFA formulation: Eq. (1-7)

$$b_k^F + b_k^R + z_k \leq 1, \ \forall k \in R^t \quad (14)$$

$$v_{MT} \geq c \cdot v_{MT,max} \quad (15)$$

where R^M is the set of all the reactions in the model. z_k, b_k^F and b_k^R are binary variables to control the activity and the flux through the ith reaction of R^M. v_{MT} is the flux through the sink reaction for the metabolic task and $v_{MT,max}$ is the maximum flux through the metabolic task as a result of applying thermodynamic flux variability analysis. c is a parameter that controls the flux requirement for the metabolic task. For this study we have chosen $c = 0.9$, to demand at least 90% of the maximum production of the metabolite for the associated metabolic task. As a result, if $z_k = 1$, then the corresponding reaction cannot carry flux.

v. Generate alternative sets using the MILP formulation (step iv). The algorithm allows to generate alternatives of minimum size ($smin$) or larger, as $smin + n$, $n = 1,2,3, \ldots$

With this formulation, the MiN contains intracellular reactions, transport reactions and boundary reactions for extracellular metabolites that are required for each metabolic task.

3.4.11 Minimal Network Enrichment Analysis (MiNEA)

The reactions that compose the MiN inherit the gene-protein-reaction rules (GPR rules) from the model. MiNEA [53] uses the GPR rules and the transcriptomics data to identify the genes that are deregulated, and it assigns a deregulation score to the overall minimal network based on the deregulation of the reactions that compose it. The algorithm then identifies the most deregulated minimal network for each task by ranking the alternative minimal networks based on their overall deregulation (see [53] for further details).

The significance of the deregulation of each MiN is tested using the multivariate Fisher's hypergeometric distribution for which the p-value for upregulated reactions is computed as it follows:

$$pvalue = \sum_{i=T^{up}}^{\min(R^{up},T)} \sum_{j=0}^{R^{down}} \frac{\binom{R^{up}}{i}\binom{R^{down}}{j}\binom{R^{no_{dereg}}}{T-i-j}}{\binom{R}{T}} \quad (16)$$

where R and T represent the total number of reactions in the model and the MiN, respectively. R^{up}, R^{down}, R^{no_dereg} denote the number of upregulated, downregulated and unregulated reactions in the model respectively, and T^{up}, T^{down}, T^{no_dereg} the number of reactions with the corresponding deregulation in the MiN.

The p-value for downregulated reactions is computed, replacing in equation (16) R^{up} with R^{down} and R^{down} with R^{up} (see [53] for further details).

Finally, MiNEA computes the percentage of significantly deregulated MiNs for each metabolic task.

3.5 Author contribution

The work in this chapter is *in preparation* to be submitted for publication with the provisory title: *Model-based data integration and minimal network enrichment analysis identifies metabolic differences across cancer types*.

All the research in this chapter, including the literature collection and processing of the cancer data, the metabolic modeling, the integration of the data into the models, the code and simulations, the writing of the manuscript as well as all the figures were performed by Maria Masid under the supervision of Prof. Vassily Hatzimanikatis.

References

1. Weinstein, I.B. and K. Case, *The history of cancer research: Introducing an AACR Centennial series.* Cancer Research, 2008. **68**(17): p. 6861-6862.
2. Hanahan, D. and R.A. Weinberg, *Hallmarks of Cancer: The Next Generation.* Cell, 2011. **144**(5): p. 646-674.
3. Pavlova, N.N. and C.B. Thompson, *The Emerging Hallmarks of Cancer Metabolism.* Cell Metab, 2016. **23**(1): p. 27-47.
4. Junttila, M.R. and F.J. de Sauvage, *Influence of tumour micro-environment heterogeneity on therapeutic response.* Nature, 2013. **501**(7467): p. 346-54.
5. Ganapathy-Kanniappan, S., *Editorial: Cancer Metabolism: Molecular Targeting and Implications for Therapy.* Front Oncol, 2017. **7**: p. 232.
6. DeBerardinis, R.J., et al., *The biology of cancer: metabolic reprogramming fuels cell growth and proliferation.* Cell Metab, 2008. **7**(1): p. 11-20.
7. Deberardinis, R.J., et al., *Brick by brick: metabolism and tumor cell growth.* Curr Opin Genet Dev, 2008. **18**(1): p. 54-61.
8. Kroemer, G. and J. Pouyssegur, *Tumor cell metabolism: cancer's Achilles' heel.* Cancer Cell, 2008. **13**(6): p. 472-82.
9. Vander Heiden, M.G. and R.J. DeBerardinis, *Understanding the Intersections between Metabolism and Cancer Biology.* Cell, 2017. **168**(4): p. 657-669.
10. Thiele, I. and B.O. Palsson, *A protocol for generating a high-quality genome-scale metabolic reconstruction.* Nature Protocols, 2010. **5**(1): p. 93-121.
11. Folger, O., et al., *Predicting selective drug targets in cancer through metabolic networks.* Molecular Systems Biology, 2011. **7**.
12. Zielinski, D.C., et al., *Systems biology analysis of drivers underlying hallmarks of cancer cell metabolism.* Scientific Reports, 2017. **7**.
13. Shlomi, T., et al., *Genome-Scale Metabolic Modeling Elucidates the Role of Proliferative Adaptation in Causing the Warburg Effect.* Plos Computational Biology, 2011. **7**(3).

14. Warburg, O., *On the metabolism of cancer cells.* Naturwissenschaften, 1924. **12**: p. 1131-1137.

15. Fenninger, L.D. and G.B. Mider, *Energy and Nitrogen Metabolism in Cancer.* Advances in Cancer Research, 1954. **2**: p. 229-251.

16. Cairns, R.A., I.S. Harris, and T.W. Mak, *Regulation of cancer cell metabolism.* Nature Reviews Cancer, 2011. **11**(2): p. 85-95.

17. Dang, C.V., *Links between metabolism and cancer.* Genes & Development, 2012. **26**(9): p. 877-890.

18. Zur, H., E. Ruppin, and T. Shlomi, *iMAT: an integrative metabolic analysis tool.* Bioinformatics, 2010. **26**(24): p. 3140-2.

19. Agren, R., et al., *Identification of anticancer drugs for hepatocellular carcinoma t hrough personalized genome-scale metabolic modeling.* Mol Syst Biol, 2014. **10**: p. 721.

20. Wang, Y., J.A. Eddy, and N.D. Price, *Reconstruction of genome-scale metabolic models for 126 human tissues using mCADRE.* BMC Syst Biol, 2012. **6**: p. 153.

21. Vlassis, N., M.P. Pacheco, and T. Sauter, *Fast reconstruction of compact context-specific metabolic network models.* PLoS Comput Biol, 2014. **10**(1): p. e1003424.

22. Jerby, L., T. Shlomi, and E. Ruppin, *Computational reconstruction of tissue-specific metabolic models: application to human liver metabolism.* Molecular Systems Biology, 2010. **6**.

23. Opdam, S., et al., *A Systematic Evaluation of Methods for Tailoring Genome-Scale Metabolic Models.* Cell Syst, 2017. **4**(3): p. 318-329 e6.

24. Fouladiha, H. and S.A. Marashi, *Biomedical applications of cell- and tissue-specific metabolic network models.* Journal of Biomedical Informatics, 2017. **68**: p. 35-49.

25. Cho, J.S., et al., *Reconstruction of context-specific genome-scale metabolic models using multiomics data to study metabolic rewiring.* Current Opinion in Systems Biology, 2019. **15**: p. 1-11.

26. Soh, K.C.a.H., V., *Constraining the flux space using thermodynamics and integration of metabolomics data.* Methods Mol Biol (Clifton, N.J.), 2014(1191): p. pp. 49-63.

27. Henry, C.S., L.J. Broadbelt, and V. Hatzimanikatis, *Thermodynamics-based metabolic flux analysis.* Biophysical Journal, 2007. **92**(5): p. 1792-1805.

28. Jankowski, M.D., et al., *Group contribution method for thermodynamic analysis of complex metabolic networks.* Biophysical Journal, 2008. **95**(3): p. 1487-1499.

29. Salvy, P., et al., *pyTFA and matTFA: a Python package and a Matlab toolbox for Thermodynamics-based Flux Analysis.* Bioinformatics, 2018. **35**(1): p. 167-169.

30. Ataman, M., et al., *redGEM: Systematic reduction and analysis of genome-scale metabolic reconstructions for development of consistent core metabolic models.* Plos Computational Biology, 2017. **13**(7).

31. Ataman, M. and V. Hatzimanikatis, *lumpGEM: Systematic generation of subnetworks and elementally balanced lumped reactions for the biosynthesis of target metabolites.* Plos Computational Biology, 2017. **13**(7).

32. Jain, M., et al., *Metabolite Profiling Identifies a Key Role for Glycine in Rapid Cancer Cell Proliferation.* Science, 2012. **336**(6084): p. 1040-1044.

33. Brunk, E., et al., *Recon3D enables a three-dimensional view of gene variation in human metabolism.* Nature Biotechnology, 2018. **36**(3): p. 272-+.

34. Navale, A.M. and A.N. Paranjape, *Glucose transporters: physiological and pathological roles.* Biophys Rev, 2016. **8**(1): p. 5-9.

35. Kilburn, D.G., M.D. Lilly, and F.C. Webb, *The energetics of mammalian cell growth.* J Cell Sci, 1969. **4**(3): p. 645-54.

36. Pandey, V., et al., *TEX-FBA: A constraint-based method for integrating gene expression, thermodynamics, and metabolomics data into genome-scale metabolic models.* bioRxiv doi: 10.1101/536235, 2019.

37. Frezza, C., *Histidine degradation boosts cancer therapy.* Nature, 2018. **559**(7715): p. 484-485.

38. Kanarek, N., et al., *Histidine catabolism is a major determinant of methotrexate sensitivity.* Nature, 2018. **559**(7715): p. 632-+.

39. Hay, N., *Reprogramming glucose metabolism in cancer: can it be exploited for cancer therapy?* Nature Reviews Cancer, 2016. **16**(10): p. 635-649.

40. Keshet, R. and A. Erez, *Arginine and the metabolic regulation of nitric oxide synthesis in cancer.* Disease Models & Mechanisms, 2018. **11**(8).

41. Tanner, J.J., S.M. Fendt, and D.F. Becker, *The Proline Cycle As a Potential Cancer Therapy Target.* Biochemistry, 2018. **57**(25): p. 3433-3444.

42. Villa, E., et al., *Cancer Cells Tune the Signaling Pathways to Empower de Novo Synthesis of Nucleotides.* Cancers, 2019. **11**(5).

43. Hanley, M.P. and D.W. Rosenberg, *One-Carbon Metabolism and Colorectal Cancer: Potential Mechanisms of Chemoprevention.* Curr Pharmacol Rep, 2015. **1**(3): p. 197-205.

44. Koseki, J., et al., *Enzymes of the one-carbon folate metabolism as anticancer targets predicted by survival rate analysis.* Scientific Reports, 2018. **8**.

45. Little, J., et al., *Colon cancer and genetic variation in folate metabolism: The clinical bottom line.* Journal of Nutrition, 2003. **133**(11): p. 3758s-3766s.

46. Lieu, E.L., et al., *Amino acids in cancer.* Experimental and Molecular Medicine, 2020. **52**(1): p. 15-30.

47. Halama, A., et al., *Metabolic signatures differentiate ovarian from colon cancer cell lines.* Journal of Translational Medicine, 2015. **13**.

48. Slattery, M.L., et al., *Dietary sugar and colon cancer.* Cancer Epidemiology Biomarkers & Prevention, 1997. **6**(9): p. 677-685.

49. He, Y., Q.D.Q.T. Li, and S.C. Guo, *Taurine Attenuates Dimethylbenz[a]anthracene-induced Breast Tumorigenesis in Rats: A Plasma Metabolomic Study.* Anticancer Research, 2016. **36**(2): p. 533-543.

50. Huang, S., et al., *Novel personalized pathway-based metabolomics models reveal key metabolic pathways for breast cancer diagnosis.* Genome Medicine, 2016. **8**.

51. UniProt, C., *UniProt: a worldwide hub of protein knowledge.* Nucleic Acids Res, 2019. **47**(D1): p. D506-D515.

52. Coordinators, N.R., *Database resources of the National Center for Biotechnology Information.* Nucleic Acids Res, 2018. **46**(D1): p. D8-D13.

53. Pandey, V. 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.* PLoS Comput Biol, 2019. **15**(4): p. e1006760.

54. Garcia-Bermudez, J., et al., *Targeting extracellular nutrient dependencies of cancer cells.* Mol Metab, 2020. **33**: p. 67-82.

55. Lane, A.N., R.M. Higashi, and T.W. Fan, *Metabolic reprogramming in tumors: Contributions of the tumor microenvironment.* Genes Dis, 2020. **7**(2): p. 185-198.

56. Eales, K.L., K.E. Hollinshead, and D.A. Tennant, *Hypoxia and metabolic adaptation of cancer cells.* Oncogenesis, 2016. **5**: p. e190.

57. Samanta, D. and G.L. Semenza, *Metabolic adaptation of cancer and immune cells mediated by hypoxia-inducible factors.* Biochim Biophys Acta Rev Cancer, 2018. **1870**(1): p. 15-22.

58. Diener, C. and O. Resendis-Antonio, *Personalized Prediction of Proliferation Rates and Metabolic Liabilities in Cancer Biopsies.* Front Physiol, 2016. **7**: p. 644.

59. Machado, D. and M. Herrgard, *Systematic evaluation of methods for integration of transcriptomic data into constraint-based models of metabolism.* PLoS Comput Biol, 2014. **10**(4): p. e1003580.

60. Hay, N., *Reprogramming glucose metabolism in cancer: can it be exploited for cancer therapy?* Nat Rev Cancer, 2016. **16**(10): p. 635-49.

61. Vander Heiden, M.G., L.C. Cantley, and C.B. Thompson, *Understanding the Warburg effect: the metabolic requirements of cell proliferation.* Science, 2009. **324**(5930): p. 1029-33.

62. Hsu, P.P. and D.M. Sabatini, *Cancer cell metabolism: Warburg and beyond.* Cell, 2008. **134**(5): p. 703-7.

63. Cluntun, A.A., et al., *Glutamine Metabolism in Cancer: Understanding the Heterogeneity.* Trends Cancer, 2017. **3**(3): p. 169-180.

64. Wise, D.R. and C.B. Thompson, *Glutamine addiction: a new therapeutic target in cancer.* Trends in Biochemical Sciences, 2010. **35**(8): p. 427-433.

65. Zhang, J., N.N. Pavlova, and C.B. Thompson, *Cancer cell metabolism: the essential role of the nonessential amino acid, glutamine.* EMBO J, 2017. **36**(10): p. 1302-1315.

66. Wallace, D.C., *Mitochondria and cancer.* Nat Rev Cancer, 2012. **12**(10): p. 685-98.

67. Ward, P.S. and C.B. Thompson, *Metabolic reprogramming: a cancer hallmark even warburg did not anticipate.* Cancer Cell, 2012. **21**(3): p. 297-308.

68. Porporato, P.E., et al., *Mitochondrial metabolism and cancer.* Cell Res, 2018. **28**(3): p. 265-280.

69. Panieri, E. and M.M. Santoro, *ROS homeostasis and metabolism: a dangerous liason in cancer cells.* Cell Death Dis, 2016. **7**(6): p. e2253.

70. Patra, K.C. and N. Hay, *The pentose phosphate pathway and cancer.* Trends Biochem Sci, 2014. **39**(8): p. 347-54.

71. Yang, M. and K.H. Vousden, *Serine and one-carbon metabolism in cancer.* Nat Rev Cancer, 2016. **16**(10): p. 650-62.

72. Mattaini, K.R., M.R. Sullivan, and M.G. Vander Heiden, *The importance of serine metabolism in cancer.* J Cell Biol, 2016. **214**(3): p. 249-57.

73. Sullivan, M.R., et al., *Increased Serine Synthesis Provides an Advantage for Tumors Arising in Tissues Where Serine Levels Are Limiting.* Cell Metab, 2019. **29**(6): p. 1410-1421 e4.

74. Labuschagne, C.F., et al., *Serine, but Not Glycine, Supports One-Carbon Metabolism and Proliferation of Cancer Cells.* Cell Reports, 2014. **7**(4): p. 1248-1258.

75. Beloribi-Djefaflia, S., S. Vasseur, and F. Guillaumond, *Lipid metabolic reprogramming in cancer cells.* Oncogenesis, 2016. **5**: p. e189.

76. Snaebjornsson, M.T., S. Janaki-Raman, and A. Schulze, *Greasing the Wheels of the Cancer Machine: The Role of Lipid Metabolism in Cancer.* Cell Metab, 2020. **31**(1): p. 62-76.

77. Santos, C.R. and A. Schulze, *Lipid metabolism in cancer.* FEBS J, 2012. **279**(15): p. 2610-23.

Appendix B

Network topology: selecting context-specific transport reactions

Table S3.1: **Transport reactions obtained with the MILP formulation**. Shadowed boxes indicate that the transport reaction is required in the corresponding cancer model.

		Breast Cancer	Colon Cancer	Ovarian Cancer
Intracellular transports	DATPtn: datp_c <=> datp_n	■	■	■
	DCTPtn: dctp_c <=> dctp_n	■	■	■
	DGTPtn: dgtp_c <=> dgtp_n	■	■	■
	DTTPtn:dttp_c <=> dttp_n	■	■	■
	FADH2tru: fadh2_c <=> fadh2_r	■	■	■
	FADtru: fad_r <=> fad_c	■	■	■
	CHSTEROLtrc: chsterol_r <=> chsterol_c	■	■	■
	FRDPtcr: frdp_c <=> frdp_r	■	■	■
	NADHtru: nadh_c <=> nadh_r	■	■	■
	NADPHtru: nadph_c <=> nadph_r	■	■	■
	NADPtru: nadp_r <=> nadp_c	■	■	■
	NADtru: nad_r <=> nad_c	■	■	■
	HMR_3953: 6pgc_c <=> 6pgc_r	■	■	■
	r0840: r5p_c <=> r5p_r	■	■	■
	r0841: ru5p_D_c <=> ru5p_D_r	■	■	■
	GLU5SAtmc: glu5sa_m <=> glu5sa_c	■	■	■
	FORtr: for_c <=> for_r	■	■	■
	ILEt5m: ile_L_c <=> ile_L_m	■		■
Extracellular transports	FUMtr: fum_e <=> fum_c	■	■	■
	MAL_Lte: mal_L_e <=> mal_L_c	■	■	■
	GUDACtr: 2 na1_e + cl_e + gudac_e <=> 2 na1_c + cl_c + gudac_c	■	■	■
	HCYSte: hcys_L_e <=> hcys_L_c	■	■	■
	ICITtr: icit_e <=> icit_c	■		
	GLCURtr: glcur_e <=> glcur_c	■	■	
	CRTNtr: crtn_e <=> crtn_c	■	■	
	GLYALDtr: na1_e + glyald_e <=> na1_c + glyald_c		■	
	AHCYSte: ahcys_e <=> ahcys_c		■	■
	XANtr: xan_e <=> xan_c		■	■
	DMGLYtr: dmgly_e <=> dmgly_c			■

Table S3.2. **Transport reactions for small molecules**

Transport Reaction Name	Formula	Transport Reaction Name	Formula
CLOHtex2	h2o_c + 2 cl_e <=> h2o_e + 2 cl_c	PIter	pi_r <=> pi_c
CO2ter	co2_c <=> co2_r	PItg	pi_g <=> pi_c
CO2tm	co2_c <=> co2_m	PItx	pi_c <=> pi_x
CO2tp	co2_c <=> co2_x	PPItr	ppi_c <=> ppi_r
H2O2t	h2o2_e <=> h2o2_c	PPItx	ppi_c <=> ppi_x
H2O2tm	h2o2_c <=> h2o2_m	SO4CLtex2	cl_c + 2 so4_e <=> cl_e + 2 so4_c
H2O2tn	h2o2_c <=> h2o2_n	SO4t4_2	2 na1_e + so4_e <=> 2 na1_c + so4_c
H2O2tp	h2o2_c <=> h2o2_x	r0838	nh4_c <=> nh4_m
H2Oter	h2o_c <=> h2o_r	r1423	pi_c <=> pi_e
H2Otg	h2o_c <=> h2o_g	r2136	na1_e + pi_e <=> pi_c + na1_c
H2Otm	h2o_c <=> h2o_m	r2521	ppi_c + pi_r <=> pi_c + ppi_r
H2Otp	h2o_c <=> h2o_x	PIt8	1.5 na1_e + pi_e <=> pi_c + 1.5 na1_c
Htg	h_g <=> h_c	PIt9	2 na1_e + pi_e <=> pi_c + 2 na1_c
Htr	h_c <=> h_r	KHte	h_c + k_e <=> h_e + k_c
Htx	h_c <=> h_x	PPItm	ppi_c <=> ppi_m
KCC2t	nh4_e + cl_e <=> nh4_c + cl_c	CLCFTRte	cl_c <=> cl_e
KCCt	k_e + cl_e <=> k_c + cl_c	r1492	k_c <=> k_e
NAt	na1_e <=> na1_c	The	h_e <=> h_c
NAt5	na1_c + nh4_e <=> nh4_c + na1_e	CO2t	co2_e <=> co2_c
NCCt	na1_e + cl_e <=> na1_c + cl_c	H2Ot	h2o_e <=> h2o_c
NH4t3r	nh4_c + h_e <=> h_c + nh4_e	NAt3_1	h_e + na1_c <=> h_c + na1_e
NH4tp	nh4_c <=> nh4_x	O2t	o2_e <=> o2_c
NKCC2t	na1_e + nh4_e + 2 cl_e <=> nh4_c + na1_c + 2 cl_c	PIt6b	h_e + pi_e <=> h_c + pi_c
NKCCt	na1_e + k_e + 2 cl_e <=> na1_c + k_c + 2 cl_c	NH4tb	nh4_e <=> nh4_c
NOt	no_e <=> no_c	PIt7	3 na1_e + pi_e <=> pi_c + 3 na1_c
O2St	o2s_c <=> o2s_e	NH4tr	nh4_c <=> nh4_r
O2Stm	o2s_c <=> o2s_m	HMR_1095	h_c <=> h_n
O2Stn	o2s_c <=> o2s_n	HMR_7700	3 na1_e + so4_e <=> 3 na1_c + so4_c
O2Stx	o2s_c <=> o2s_x	HMR_9590	2 h_e + so4_e <=> 2 h_c + so4_c
O2ter	o2_c <=> o2_r	H2O2itr	h2o2_r <=> h2o2_c
O2tm	o2_c <=> o2_m	Htmi	h_i <=> h_m
O2tn	o2_c <=> o2_n		
O2tp	o2_c <=> o2_x		

Subsystems deregulation

Table S3.3. **Percentage of the deregulated reactions for each subsystem.** Upregulated subsystems are represented in red and positive numbers, and downregulated subsystems in green and negative numbers

	Transcriptomics Data			Consistent with the Network		
	Ovarian	Colon	Breast	Ovarian	Colon	Breast
Alanine and aspartate metabolism	50	50	50	50	50	50
Aminosugar metabolism	100	100	100	100	100	100
Arginine and proline metabolism	22.2	25	22.2	22.2	25	22.2
Beta-Alanine metabolism	0	22.2	11.1	0	22.2	11.1
Bile acid synthesis	0	9.09	9.09	0	9.09	9.09
Cholesterol metabolism	2.44	9.76	4.88	2.44	9.76	4.88
Citric acid cycle	60	75	65	60	75	65
Oxidative Phosphorylation	33.3	55.6	55.6	33.3	55.6	55.6
	0	0	0	-22.2	-22.2	-22.2
Fatty acid oxidation	0	28.6	14.3	0	28.6	14.3
Fatty acid synthesis	10	70	10	10	70	10
Folate metabolism	27.3	72.7	27.3	27.3	72.7	27.3
Galactose metabolism	0	33.3	33.3	0	33.3	33.3
Glutamate metabolism	18.8	37.5	18.8	18.8	37.5	18.8
Glutathione metabolism	28.6	57.1	28.6	28.6	57.1	28.6
Glycerophospholipid metabolism	2.56	7.69	5.13	2.56	7.69	5.13
	0	-2.56	-2.56	0	-2.56	-2.56
Glycine, serine, alanine, and threonine metabolism	13.3	8.89	13.3	13.3	8.89	13.3
	0	-6.67	0	0	-13.3	0
Glycolysis/gluconeogenesis	44.7	57.9	52.6	44.7	57.9	52.6
	-2.63	0	-2.63	-2.63	0	-2.63
Glyoxylate and dicarboxylate metabolism	28.6	28.6	28.6	28.6	28.6	28.6
Histidine metabolism	-42.9	-42.9	-42.9	-42.9	-42.9	-42.9
Lysine metabolism	-12.5	-12.5	-12.5	-12.5	-12.5	-12.5
Methionine and cysteine metabolism	0	11.1	11.1	0	11.1	11.1
Miscellaneous	25	37.5	37.5	25	37.5	37.5
NAD metabolism	20	30	20	20	30	20
Nucleotide interconversion	42	63.8	56.5	42	63.8	56.5
	-2.9	-2.9	0	-2.9	-2.9	0
Nucleotide metabolism	100	100	100	100	100	100
Nucleotide salvage pathway	0	100	100	0	100	100

Pentose phosphate pathway	28.6	45.7	34.3	28.6	45.7	34.3
Phenylalanine metabolism	0	-25	0	0	-25	0
Purine catabolism	15.4	30.8	23.1	15.4	30.8	23.1
Purine synthesis	6.67	66.7	6.67	6.67	66.7	6.67
Pyrimidine catabolism	-7.14	21.4	21.4	-7.14	21.4	21.4
	0	-7.14	-7.14	0	-7.14	-7.14
Pyrimidine synthesis	10.5	10.5	15.8	10.5	10.5	15.8
Pyruvate metabolism	23.5	41.2	23.5	23.5	41.2	23.5
ROS detoxification	-5.88	-11.8	-11.8	-5.88	-11.8	-11.8
	14.3	14.3	14.3	14.3	14.3	14.3
Starch and sucrose metabolism	0	100	0	0	100	0
Tetrahydrobiopterin metabolism	0	0	0	50	0	0
	0	0	0	-50	0	0
Transport, extracellular	19.3	19.8	16.3	19.3	19.9	16.5
	-5.56	-5.27	-13.9	-5.56	-5.36	-14.1
Transport, mitochondrial	3.02	17.6	16.1	3.02	18.1	16.1
	-11.1	-11.1	-11.1	-11.1	-11.1	-11.1
Transport, peroxisomal	0	2.13	2.13	0	2.13	2.13
Triacylglycerol synthesis	0	33.3	0	0	33.3	0
Tryptophan metabolism	8	8	8	8	8	8
	-4	-4	-4	-4	-4	-4
Tyrosine metabolism	6.25	6.25	6.25	6.25	12.5	12.5
	-6.25	-25	-12.5	-6.25	-25	-18.8
Urea cycle	10.3	17.9	12.8	10.3	17.9	12.8
	-2.56	-5.13	-5.13	-2.56	-5.13	-5.13
Valine, leucine, and isoleucine metabolism	6.9	6.9	20.7	6.9	6.9	20.7
Vitamin B2 metabolism	0	50	50	0	50	50

Cancer type specific data

Table S3.4. **Cell lines associated to each type of cancer**. Cell lines clustered based on the cancer types

Cancer	Cell lines
Leukemia	CCRF-CEM, HL-60(TB), K-562, MOLT-4, RPMI-8226, SR
Lung	A549/ATCC, EKVX, HOP-62, HOP-92, NCI-H226, NCI-H23, NCI-H322M, NCI-H460, NCI-H522
Colon	COLO 205, HCC-2998, HCT-116, HCT-15, HT29, KM12, SW-620
CNS	SF-268, SF-295, SF-539, SNB-19, SNB-75, U251
Melanoma	LOX IMVI, MALME-3M, M14, MDA-MB-435, SK-MEL-2, SK-MEL-28, SK-MEL-5, UACC-257, UACC-62
Ovarian	IGR-OV1, OVCAR-3, OVCAR-4, OVCAR-5, OVCAR-8, NCI/ADR-RES, SK-OV-3
Renal	786-0, A498, ACHN, CAKI-1, RXF 393, SN12C, TK-10, UO-31
Prostate	PC-3, DU-145
Breast	MCF7, MDA-MB-231/ATCC, MDA-MB-468, HS 578T, BT-549, T-47D

Thermodynamic and expression flux variability analysis

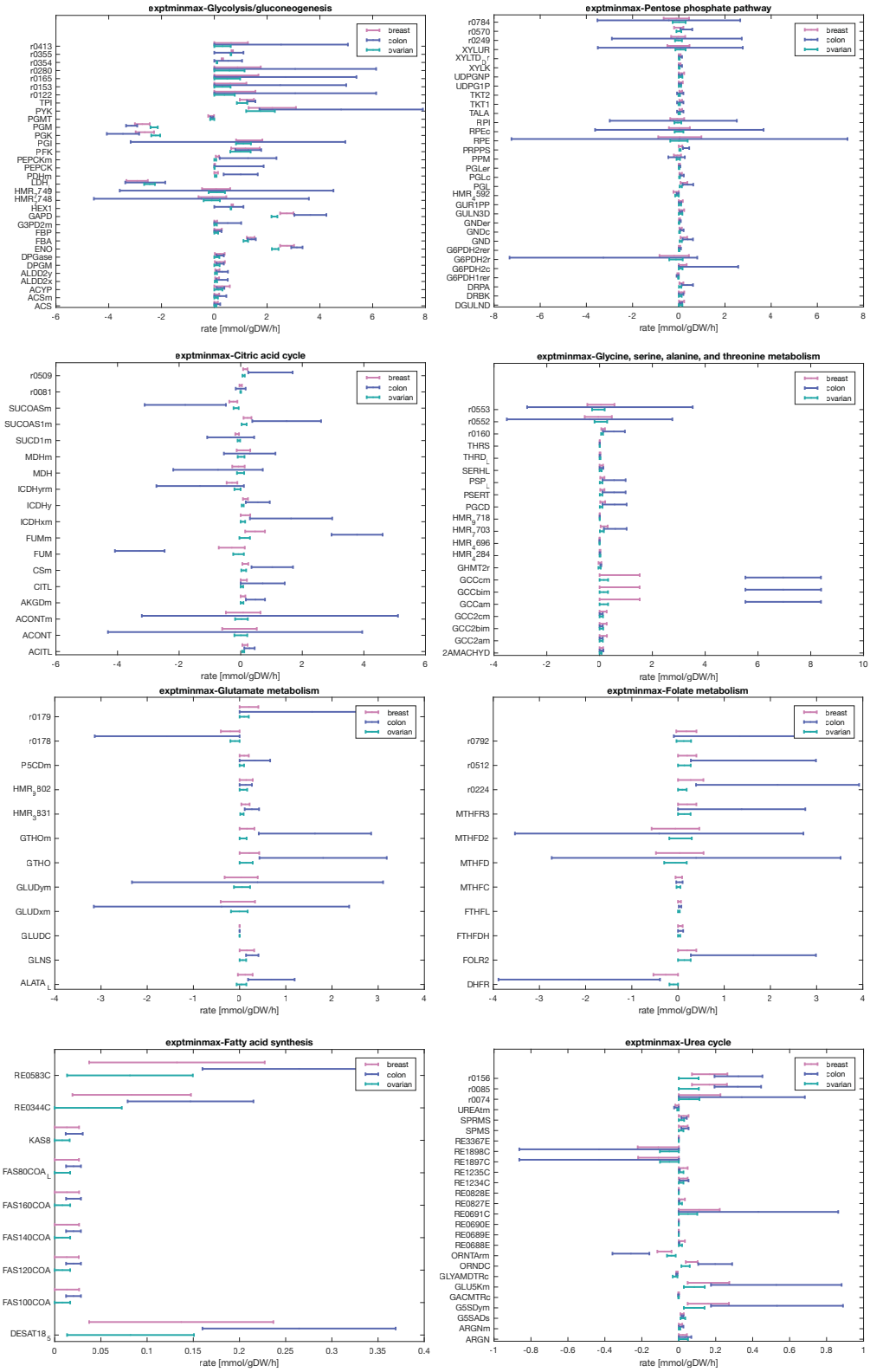


Figure S3.1. **Thermodynamic-and-expression flux variability analysis**.

Chapter 4 Integrating signaling and metabolic pathways to analyze the function of the transcription factor MYC in breast cancer

In this chapter, we develop a novel method (CONSIGN) to contextualize signaling networks to be consistent with *omics* data for a specific cell type and physiology. Moreover, we present an approach to integrate signaling and metabolic models to study interactions between both biological networks. The method here presented has been developed in collaboration with Dr. V. Pandey. A manuscript with the content of this chapter is *in preparation* to be published.

4.1 Introduction

Cells communicate with chemical signals such as proteins or other molecules, known as ligands. In that process, a cell secretes signals into the extracellular space, and the target cell, which has the right receptor for those specific signals, receives the message. The binding of ligand and receptor induces an intracellular signal sequence that causes a change in the cell and a physiological response [1, 2]. During the last decades, thousands of signaling pathways have been characterized by several processes occurring in a variety of cells under different conditions. This large amount of information is available in databases such as Reactome [3], KEGG [4], Pathway Interaction

Database [5], NetPath [6], and ACSN [7], among others. Since then, systems analysis approaches have been developed to study protein interactions and to identify alterations at the signaling level associated with several diseases, including cancer, Alzheimer's, diabetes, cardiovascular diseases, and infectious diseases [8-13].

Mathematical and computational methods have been developed for the study of signaling networks ranging from continuous models using ordinary differential equation systems to Boolean models using logic rules [14-18]. These last ones consider a discrete system with two states for the proteins in the network, active when the protein is present and triggers down the signal or inactive when the protein is absent. Boolean models rely on the structure of the network and overcome the lack of kinetic information required by the continuous models. Boolean models efficiently simulate the propagation of the signals in large scale networks in a qualitative form [19-24]. Such models have been used to integrate proteomics and transcriptomics data [25-28], to predict the effects of targeted therapies [29, 30], to develop signaling models for personalized treatments [31, 32], to model signaling alterations in diseases [33, 34], and to understand how activation of upstream pathways affect the downstream processes [27, 35].

One of the processes occurring downstream the signaling pathways is metabolism. The expression of metabolic genes is regulated by transcription factors, and their activation depends on the activity of specific signaling pathways. For example, *MYC* regulates a diversity of intracellular and extracellular processes required for cell proliferation, growth, differentiation, and death [36]. Despite its wide variety of physiological functions, *MYC* is mostly known for the role it plays in the development of cancer [37], and it represents an exciting candidate for targeted cancer therapy [38-40].

Although previous analyses have focused on the study of signaling and metabolic pathways independently, currently, there is a rising interest to develop methods that integrate both systems allowing to investigate the regulatory effects that they perform on each other. Metabolism is normally modeled using constraint-based metabolic models that impose quasi-steady-state, stoichiometric, and thermodynamic constraints to identify the flexibility of the network and the activity of the metabolic pathways. In the case of human metabolism, the most comprehensive genome-scale models are Recon 3D [41] and Human1 [42]. Several studies have recently developed methods to incorporate regulatory constraints within constraint-based metabolic models, including

rFBA [43], SR-FBA [44], iFBA [45], PROM [46], iDREAM [47], FlexFlux [48], and TRFBA [49]. Although these methods account for the regulatory constraints, they do not simultaneously simulate the signaling and metabolic networks.

In this study, we present a novel approach to study signaling interactions through the contextualization of signaling networks extracted from a knowledge database. We developed a method named contextualization of signaling networks (CONSIGN) that translates a given signaling network into a set of Boolean rules and further into a set of linear equations that are formulated into a mixed-integer linear programming optimization problem. CONSIGN allows to integrate transcriptomics and proteomics data to generate a context-specific signaling network. Then it identifies the signaling species that are maximally consistent with the experimental data, that is, the maximum number of signaling species that can be simultaneously constrained in agreement with the experimental data. The method is illustrated using a small pathway from the MYC signaling network and then applied to the overall upstream signaling pathways related to MYC. Furthermore, we present a novel workflow to integrate signaling and metabolic networks, which allows us to investigate regulatory interactions. In this work, we start with a generic signaling network map from Reactome, and by using our approach combined with transcriptomics data, we are able to identify signaling components that are relevant to the particular type of cancer. Specifically, we identify active signaling events and components in breast cancer cells, integrating transcriptomics data from the breast cancer NCI60 cell lines [50] into a signal transduction model for the transcription factor. Using this workflow, we created an integrated model for the MYC signaling model and the metabolic model for breast cancer. The metabolic model was generated by integrating metabolomics and transcriptomics data from the breast cancer NCI60 cell lines into a reduced version of the human metabolic genome-scale model Recon 3D following the workflow defined in Chapter 3. The MYC-breast cancer model is used to integrate transcriptomics and analyze the consistency of the data with both biological networks.

4.2 Results

4.2.1 Method overview

The workflow here developed integrates context-specific signaling and metabolic networks (Figure 4.1). For a set of signaling species, the upstream or downstream signaling networks are extracted from the signaling database Reactome [3]. The workflow includes a novel method named contextualization of signaling networks (CONSIGN) to generate context-specific signaling models. CONSIGN translates the signaling network into a set of Boolean rules, which are then converted to linear equations forming a mixed-integer linear programming (MILP) problem. Then CONSIGN identifies the maximum number of signaling components whose activity in the network is consistent with transcriptomics or proteomics data for a given context. Furthermore, the gene-protein-reaction (GPR) rules in the metabolic network are formulated as Boolean rules and integrated as constraints into the thermodynamic-flux balanced metabolic problem [51]. Then, we identify the transcription factors that promote metabolic genes from the metabolic network. The MILP problem formulation of both the signaling network and the metabolic network are then integrated using Boolean rules to connect the transcription factors to the genes in the metabolic network based on the regulation (activation or repression). The activation or repression of the genes is related to the activation or inactivation of the enzymes and, thus, to the fluxes of the metabolic reactions in the network. The combined model, including the signaling and metabolic networks, is then used to integrate expression data and analyze the possible states of the networks.

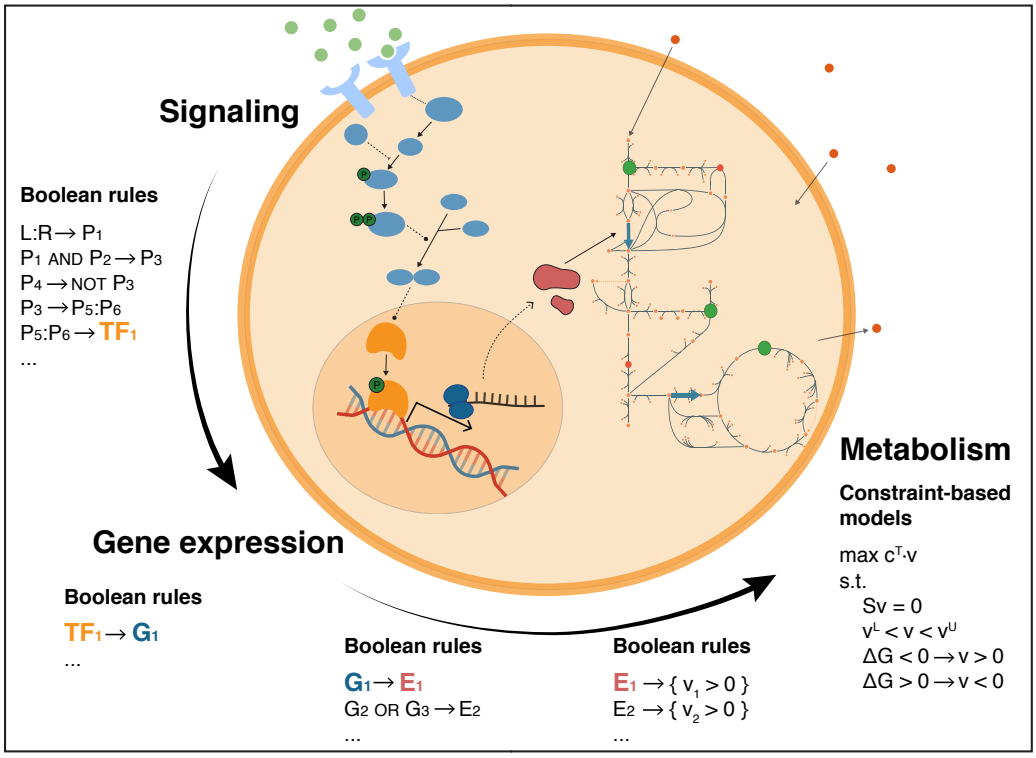


Figure 4.1. **Method overview.** Signaling events are modeled using Boolean rules that represent protein activation, complex formation, reaction activation, and reaction inhibition. Upon binding of the ligand and receptor (L:R), the signaling proteins activate (P_i) transmitting the signal. When a protein or a complex (P_5:P_6) activates a transcription factor (TF_1), we assume that the corresponding regulated gene (G_1) is active. We use Boolean rules to model the regulation of the expression of a gene by a transcription factor. Then, the active gene transcribes to the corresponding enzyme (E_i), which is produced to an adequate degree to generate flux through the corresponding reaction (v_i). The gene-enzyme-reaction relationship is modeled by Boolean rules following the GPR rules described in the metabolic model. Finally, the reactions follow the stoichiometric and thermodynamic constraints defined in the constraint-based formulation of the metabolic model.

4.2.2 The transcription factor MYC and its relation to cancer

The transcription factor MYC regulates a diversity of intracellular and extracellular transcriptional programs required for the correct proliferation of cells. These programs include the regulation of cell growth, cell cycle, metabolism, protein biosynthesis, microRNAs expression, invasion, and angiogenesis, as well as protective mechanisms including growth arrest and apoptosis [38, 52-54]. The deregulation of the *MYC* gene has been associated with cancer development. *MYC* overexpression induces changes in the expression of several genes to increase cellular proliferation, driving at the same time metabolic changes to support the increased demand for proteins, lipids, and nucleic

acids. This interplay between the alterations in the upstream and downstream signaling pathways for MYC and the effects induced at the metabolic level highlights the importance of the oncogene MYC in shaping the cancer phenotype [55].

4.2.3 A small signaling pathway for MYC

As a first case study, we extracted from Reactome a pathway of ten reactions and 20 species that represents a negative feedback loop to control the expression of *MYC* (Figure 4.2 A). The pathway starts with the transcription of the gene *MAPKAPK5* upon binding of MYC to the MAPKAPK5 promoter in the nucleoplasm of the cell. Then p-S189 MAPK6 and p-S186 MAPK4 (represented in the network as p-S MAPK6,4) bind to MAPKAPK5 to form the complex p-S MAPK6,4:MAKPAK5. The activated MAPK6 and MAPK4 promote the phosphorylation of MAPKAPK5, represented in the network as p-S MAPK6,4:P-T182 MAPKAPK5. The proteins of the activated complex are then redistributed to the cytosol. In the next step, the activated MAPKAPK5 phosphorylates FOXO3, promoting its activation and translocation to the nucleus. In the nucleus, the phosphorylated FOXO3 binds to the *miR-34B and C* genes, promoting the expression of the microRNAs. Finally, miR-34 microRNAs bind and cause the degradation of MYC mRNA, negatively regulating the translation of MYC mRNA and thus decreasing the level of MYC protein.

We developed a method named contextualization of signaling networks (CONSIGN, Materials and Methods) to build a Boolean model from a reconstructed signaling pathway and to identify the signaling species whose activity is consistent with experimental measurements. The method starts by converting the signaling events into logic rules. In these rules, the substrates and the activators activate the reaction, and then, the activation of the reaction results in the formation of the products. For example, in the reconstructed negative feedback pathway for MYC, the first reaction of the pathway, named R-HSA-5687115 in Reactome, requires the *MAPKAPK5* gene and the complex MAPKAPK5 gene:MYC as substrate and activator respectively, and it produces the protein MAPKAPK5. The logic rules associated with this reaction are the following:

$$MAPKAPK5\ gene \text{ AND } MAPKAPK5gene{:}\ MYC \rightarrow reaction_{R-HSA-5687115}$$
$$reaction_{R-HSA-5687115} \rightarrow MAPKAPK5$$

The states of the species, including proteins, complexes, and reactions, are represented with binary variables and the following algebraic linear equations (Materials and Methods) are used to describe the AND rule:

$$2x + 2y - 4z \geq -1,$$
$$2x + 2y - 4z \leq 3,$$

where $x, y,$ and z represent *MAPKAPK5* gene, MAPKAPK5 gene:MYC, and the reaction R-HSA-5687115, respectively.

Then the following equation represents the formation of the product,

$$z - p = 0,$$

where z and p are binary variables that represent the states of the reaction R-HSA-5687115 and the product MAPKAPK5, respectively.

As another example, the last reaction of the pathway, R-HSA-5687115, describes the inhibition of the translation of MYC mRNA into MYC by miR-34B,C RISC. The logic rules associated with this reaction are the following:

$$MYC\ mRNA \text{ AND not } miR-34B, C\ RISC \rightarrow reaction_{R-HSA-5687113}$$
$$reaction_{R-HSA-5687113} \rightarrow MYC$$

In this case, binary variables are created for the species, and an additional binary variable is necessary to represent `not`(miR-34B,C RISC). The corresponding set of algebraic equations in the model are the following:

$$y + y^{not} = 1,$$
$$2x + 2y^{not} - 4z \geq -1,$$
$$2x + 2y^{not} - 4z \leq 3,$$
$$z - p = 0,$$

where x represents MYC mRNA, y represents miR-34B,C RISC, y^{not} represents `not`(miR-34B,C RISC), z represents the reaction R-HSA-5687113, and p represents the product MYC.

Similarly, we generated rules for the ten reactions in the pathway, and we formulated a mixed-integer linear programming (MILP) optimization problem that allows creating a context-specific signaling network by integrating transcriptomics and proteomics data.

The context-specific network allows hypothesizing possible states of the network that can explain the experimental data.

In this study, we used transcriptomics data from the NCI60 cell lines for breast cancer. We discretized the transcriptomics data into three levels: high, medium, and low, by analyzing the expression values across samples (Materials and Methods). We considered single proteins and genes as observable species, and we defined their state as active if they were highly expressed and as inactive if they were lowly expressed in the discretized transcriptomics data. In addition, we considered that chemical compounds would also be present in the cell. In this case, we defined as active only the chemical compounds always acting as substrates in the reactions. And their usage is determined depending on the activity of the reaction by solving the MILP problem to maximize consistency with the data.

For the reconstructed MYC network, we identified ten observable states, corresponding to two chemical compounds (ATP and ADP) and eight proteins, including MAPKAPK5, FOXO3, and MYC (Figure 4.2 A). From the breast cancer NCI60 cell lines, two of the observable states were active, namely, FOXO3 and MYC. Notice that in the absence of phosphoproteomics data, we do not assign any state to the phosphorylated proteins, as it is the case in this network for phosphorylated FOXO3. We defined only the state for the unphosphorylated protein, and the network will assign a specific state to the phosphorylated protein based on the constraints imposed to the other proteins from the observed transcriptomics data.

As a result of the optimization problem generated with CONSIGN the network can simulate the flux of information to represent the states of the proteins FOXO3 and MYC, and the metabolite ATP consistently with the data, that is, the network can express these three proteins as active. Here, we show two possible alternative states of the network (Figure 4.2 B) which are two instances of the network with different patterns of activation that can equally explain the observed data. In the first case, the cascade initiated with the transcription of *MAPKAPK5* is active, promoting the phosphorylation of FOXO3. However, the phosphorylated FOXO3 cannot bind to the promotor preventing the transcription of the microRNAs. Consequently, MYC mRNA can be translated to MYC. In a second case, MYC did not bind to the MAPKAPK5 promotor preventing the transcription of MAPKAPK5. As a result, the cascade is inactive, and the phosphorylation

of FOXO3 does not occur. In this second case, MYC mRNA can be translated to MYC protein.

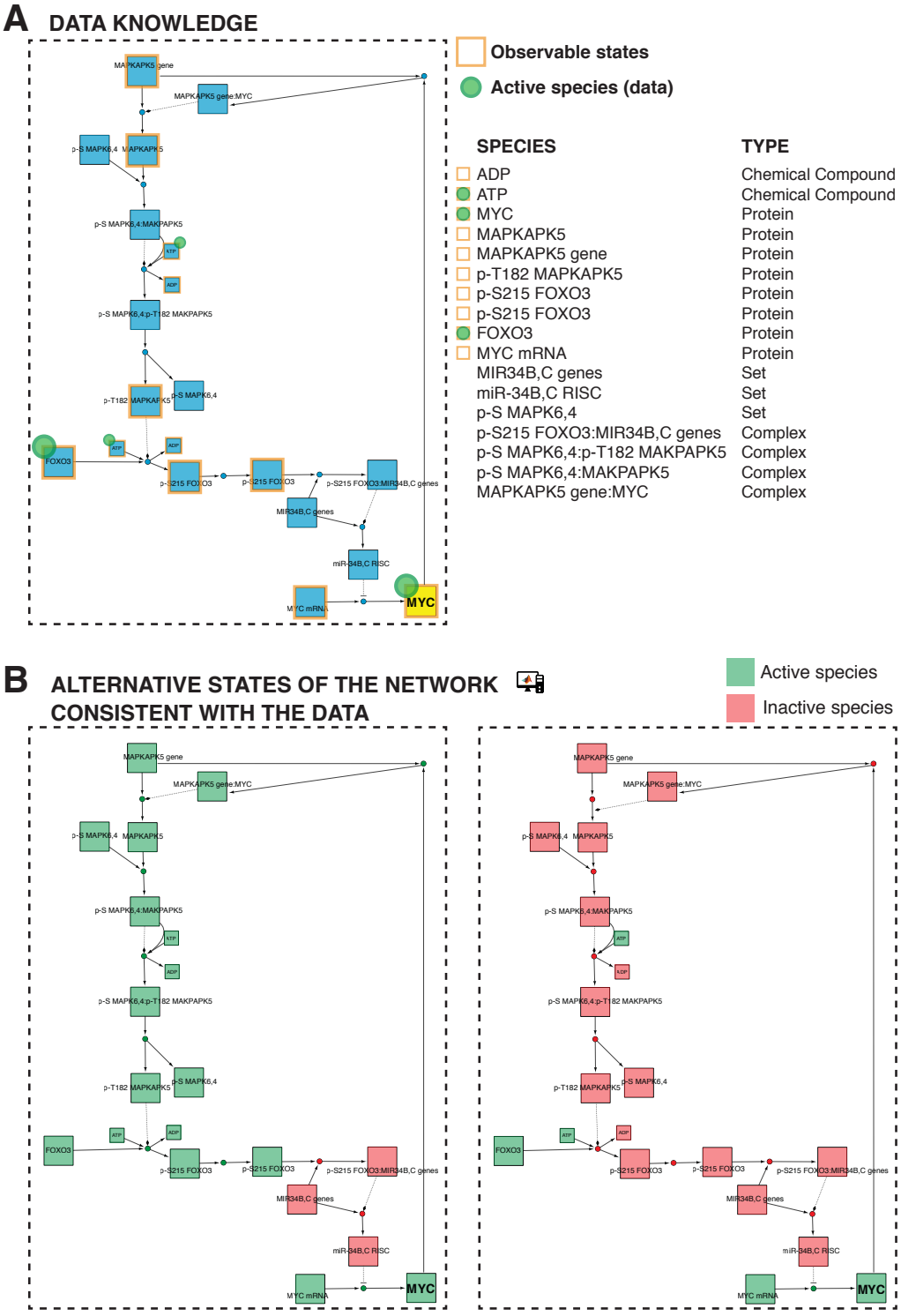


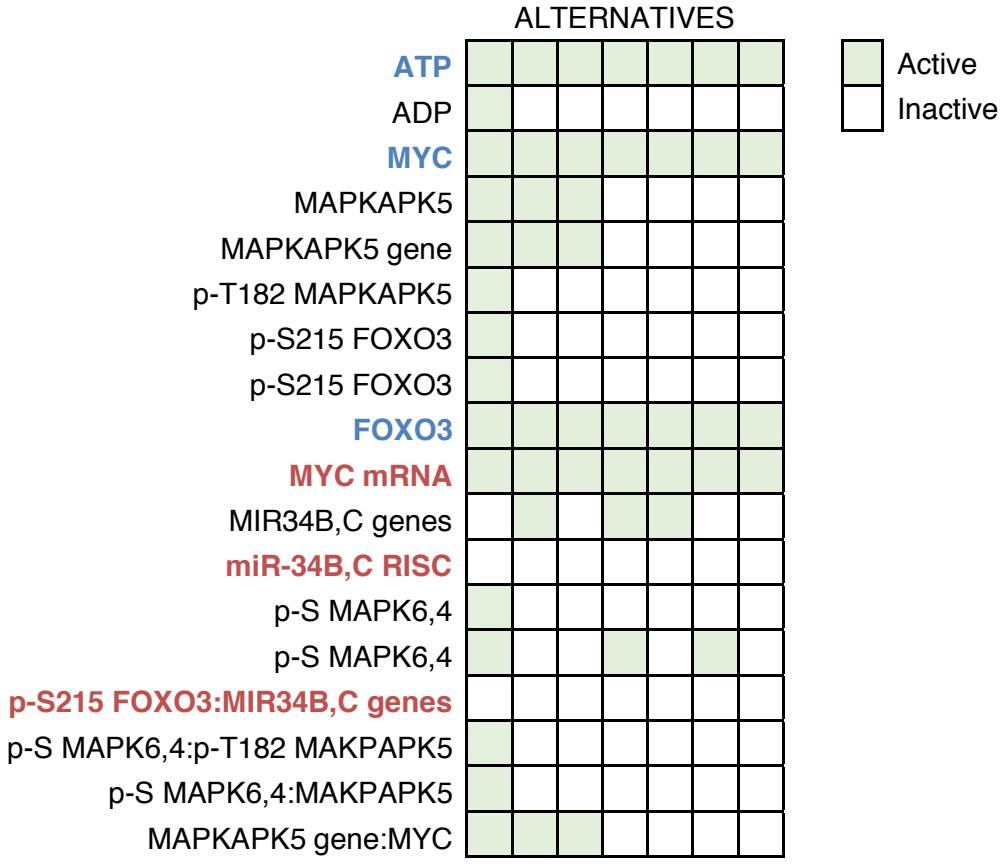
Figure 4.2. **MYC signaling pathway**. (A) A branch of the MYC signaling network upstream ten reactions from MYC. Proteins and Genes are considered in the workflow as observable states. We integrate data for the observable states if they are part of the dataset. (B) Two alternative

states of the network consistent with the data obtained as solutions of the MILP to maximize consistency after data integration.

In total, there are seven activation patterns of the network that allow maximum consistency with the experimental data, that is, alternative states of the signaling network that can explain the active states of the two proteins, FOXO3 and MYC (Table 4.1).

We observed that the state of six species, including ATP, MYC, MYC mRNA, miR-34B, C RISC, and p-S215 FOXO3:MIR34B,C genes, is always consistent across alternatives. Therefore, using the network and CONSIGN, we can infer the behavior of species that are not part of the initial data, but whose state is defined based on the consistency of the integrated data.

Table 4.1. **Alternative states of the negative feedback loop network for MYC**. Seven alternative states of the network that are consistent with the three species active in the transcriptomics data (blue). The state of another three species can be inferred from the integration of the transcriptomics data (orange).



The systematic analysis of the alternative network states that can explain the data indicate the flexibility of the pathway to propagate the signal from the inputs of the pathway, such as the activation of a receptor, to the target protein. Moreover, we have the opportunity to understand the possible flows of information that are consistent with the data as the signal propagates through the network. Further curation using literature knowledge and discussion with the experts can help to characterize the results and to further investigate which are the most biologically relevant states of the network among the alternatives proposed by our method.

4.2.4 Interactions of the signaling pathway and metabolism

Transcription factors bind to specific regions of DNA to initiate and regulate the transcription of target genes to mRNA, promoting, or repressing the expression of the proteins. Therefore, transcription factors are responsible for the gene expression pattern in the different cell types and cell states. A set of transcription factors target metabolic genes, regulating the metabolic phenotype of cells, as promoting or repressing a metabolic gene will impact the availability of the enzyme, and consequently, the activity of the metabolic pathway.

Considering the state of the transcription factors upon activation or repression as the output of the signaling network and as the input for the regulation of the metabolic genes, we can connect both the signaling and the metabolic networks, generating an integrated network to analyze not only the signal transduction but also the metabolic response to that signal.

In this work, we defined the following novel workflow to connect signaling and metabolic networks: (i) translate the gene-protein-reaction (GPR) rules into constraints using the TIGER toolbox [56] and integrate these constraints in the TFA formulation of the metabolic model (Materials and Methods); (ii) identify the transcription factors that target metabolic genes using the TRRUST database [57]; (iii) generate rules that describe the interactions of the transcription factors with the metabolic genes (Materials and Methods) and translate these rules into algebraic equations that can be integrated as constraints in the models; (iv) build an optimization problem including the signaling model, the metabolic model resulting from step (i) and the constraints generated in step (iii) that describe the regulation of the metabolic genes by transcription factors.

We connected the ten reactions MYC pathway to the metabolic model for breast cancer generated in Chapter 3. Based on the genes that are part of the human genome-scale model Recon 3D [41], MYC regulates the expression of 12 metabolic genes, including activation of *ASS1*, *CD38*, *FUT3*, *LDHA*, *ODC1*, *PRDX3*, *ST3GAL1*, *ST3GAL3*, and *ST3GAL4*, and repression of *BCAT1*, *CHKA*, and *PRODH* (Figure 4.3). The breast cancer metabolic model contains five of these genes, namely, *LDHA*, *ODC1*, *PRDX3*, *CHKA,* and *PRODH*. These genes encode for the enzymes lactate dehydrogenase (LDH_L) which converts pyruvate into lactate, glyoxylate oxidase (GLXO1) that interconverts glyoxylate and oxaloacetate, ornithine decarboxylase (ORNDC and HMR_4422) which transforms ornithine into putrescine, one enzyme in cytosol and the another one in the extracellular space, glutathione peroxidase mitochondria (GTHPm), which oxidizes glutathione in mitochondria, choline kinase (CHOLK) which transforms choline to choline-phosphate, and proline dehydrogenase (PROD2m) that converts the amino acid L-proline into pyrroline-5-carboxylate.

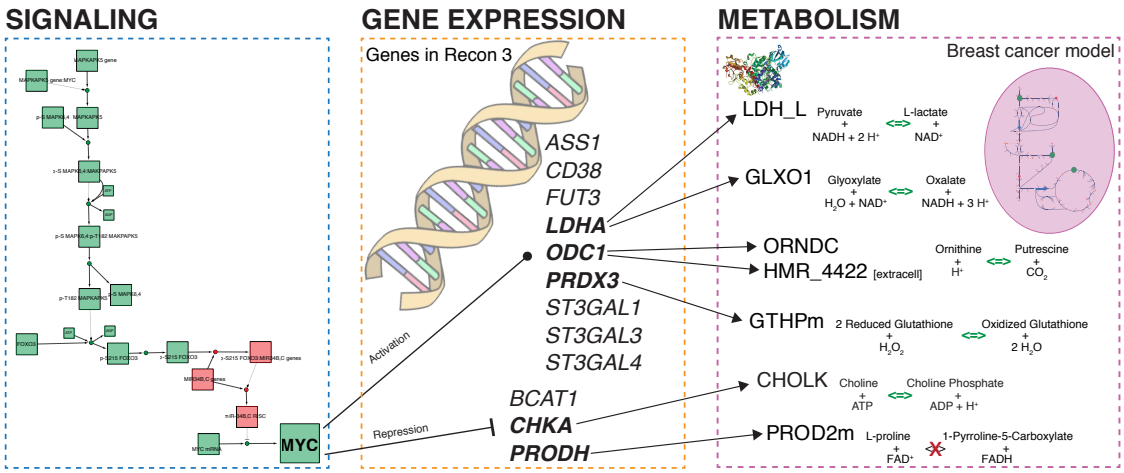


Figure 4.3. **MYC subnetwork connected to the metabolic network**. MYC is a transcription factor that regulates 12 metabolic genes present in Recon 3D. Specifically, it promotes the expression of argininosuccinate synthase (*ASS1*), ADP-ribosyl cyclase (*CD38*), Galactoside 3(4)-L-fucosyltransferase (*FUT3*), L-lactate dehydrogenase A chain (*LDHA*), Ornithine decarboxylase (*ODC1*), Thioredoxin-dependent peroxide reductase (*PRDX3*), CMP-N-acetylneuraminate-beta-galactosamide-alpha-2,3-sialyltransferase 1 (*ST3GAL1*), CMP-N-acetylneuraminate-beta-1,4-galactoside alpha-2,3-sialyltransferase (*ST3GAL3*), CMP-N-acetylneuraminate-beta-galactosamide-alpha-2,3-sialyltransferase 4 (*ST3GAL4*), and it inhibits the expression of Branched-chain-amino-acid aminotransferase (*BCAT1*), Choline kinase alpha (*CHKA*), and Proline dehydrogenase 1 (*PRODH*). Five of those genes encode enzymes that catalyze reactions that are part of the breast cancer-specific metabolic model used in this study, namely, L-lactate dehydrogenase (LDH_L), Glyoxylate oxidase (GLXO1), Ornithine Decarboxylase (ORNDC and HMR_4422), Choline kinase (CHOLK) and Proline dehydrogenase (PROD2m).

We used equation (1) to model that MYC promotes the transcription of the metabolic genes *LDHA*, *ODC1* and *PRDX3*, and equation (1) with the `not` operator to model the inhibition of *CHKA* and *PRODH* by MYC. Then, the genes are directly connected to the enzymes by the constraints derived from the GPR rules defined in the metabolic model. Finally, the enzymes constrain the fluxes through the corresponding reactions. Simulating growth with the integrated model, we observe that if MYC is activated, the expression of the genes that are promoted by MYC is also activated, and the reactions catalyzed by the enzymes coded by these genes are operative. On the other hand, when MYC is active, the transcription of *CHKA* and *PRODH* is inhibited, and this repression is reflected in the metabolic network. When the enzyme PROD2m is repressed, the reaction cannot carry flux. In the case of the enzyme CHOLK, we still observe flux through the reaction because CHOLK is encoded in the GPR of the metabolic model not only by *CHKA* but also by *CHKB*. MYC only represses one of these two genes; thus, the enzyme is still transcribed by the other gene, and the reaction can carry flux.

Having integrated both models in the same optimization problem, we now have the opportunity to incorporate the transcriptomics data at both signaling and metabolic levels. We first mapped the discretized transcriptomics data from the breast cancer NCI60 cell lines to the proteins and genes in the integrated model, and we identified a total of 244 active proteins and genes and 134 inactive genes. Specifically, two signaling active proteins MYC, and FOXO3, 242 active genes, and 134 inactive genes. Additionally, we considered ATP as active in the signaling network, and we required the maximum production of biomass in the metabolic network. In this case, we applied CONSIGN to integrate the expression data and maximize the consistency of the network states with the data considering not only the signaling species but also the expression of the metabolic genes. The results reveal that 378 of the 380 states of the proteins and genes reported in the transcriptomics data are consistent with the network, that is, their values in the integrated model can be defined according to their regulated states in the data. The two inconsistent genes encode for two subunits of the enzyme cytochrome c oxidase, namely, *COX7B2* and *COX8C*. These two genes are reported as downregulated in the transcriptomics data; however, the network requires the activity of the corresponding enzyme. The enzyme cytochrome c oxidase is the last enzyme in the mitochondrial electron transport chain, and its activity is essential to simulate growth.

4.2.5 Upstream signaling pathway for MYC and metabolism of breast cancer

Aiming to identify all the interactions between the signaling pathways related to *MYC* expression in human breast cancer metabolism, we generated a signaling network containing all the upstream pathways for the transcription factor MYC. The upstream MYC signaling network was reconstructed from Reactome, and it is composed of 3838 reactions and 6642 species, including 1963 activators, 166 inhibitors, and 30 receptors. In this network, MYC protein formation is promoted by NOTCH1 coactivator complex and by NOTCH1 PEST domain mutants coactivator complex, and it is repressed by the microRNAs 38B and C, and by a complex formed by RBL1, E2F4/5, DP1/2, p-RSMADS and, TIE (a TGF-beta inhibitory element).

Although the network was generated as the upstream signaling network for the transcription factor MYC, it contains 29 additional transcription factors that regulate a total of 71 metabolic genes present in the breast cancer model (Table 4.2). In this case, we connected the generated MYC signaling network to the metabolic breast cancer model by considering the regulation of the 71 metabolic genes by the corresponding transcription factors. To this end, we generated rules that describe the interactions of genes and transcription factors (Materials and Methods). As an example, the transcription of the gene *LDHA* that codes for the enzyme lactose dehydrogenase is promoted by five transcription factors, namely, HIF1A, HSF1, JUN, MYC, and SP1. The rule created for this gene is the following:

$$HIF1A \text{ OR } HSF1 \text{ OR } JUN \text{ OR } MYC \text{ OR } SP1 \leftrightarrow LDHA$$

We modeled the gene transcription repression using the `not` operator, as in the case of the transcription of the gene *SLC2A1* that codes for a glucose transporter. *SLC2A1* is promoted by HIF1A and repressed by ATM and TP53. For this gene, the following rule represents its regulation:

$$HIF1A \text{ AND not } ATM \text{ AND not } TP53 \leftrightarrow SLCA1$$

The rules are translated into linear equations and integrated as constraints in the optimization problem of the signaling and metabolic models combined (Materials and Methods).

Table 4.2. **Transcription factors and metabolic genes**. Transcription factors (TFs) that regulate metabolic genes present in the breast cancer specific model. The corresponding regulation is specified as activation [A] or repression [I]. The two genes used as examples in the text are highlighted in blue and orange in the table.

TFs	Genes [Regulation]
APC	*ODC1* [I]
APEX1	*SLC5A5* [A]
ATF2	*PCK1* [A]
ATM	***SLC2A1*** **[I]**
CEBPB	*GOT1* [A], *PCK2* [A], *SLC19A1* [A], *SLC5A8* [A]
CEBPD	*SOD1* [A]
CTNNB1	*PLD1* [A]
EGR1	*SLC4A2* [A], *SLC9A3* [A], *SOD1* [A]
HDAC1	*GAD1* [I]
HDAC2	*PRDX2* [I]
HIF1A	*ACE2* [I], *ALDOA* [A], *CA9* [A], ***LDHA*** **[A]**, *NT5E* [A], *PGK1* [A], *SDHB* [I], *SLC29A1* [I], ***SLC2A1*** **[A]**
HMGA1	*SLC2A3* [A]
HSF1	***LDHA*** **[A]**
JUN	***LDHA*** **[A]**, *MAT2A* [A]
KLF4	*HDC* [I], *ODC1* [I]
KLF5	*FASN* [A]
MITF	*ACP5* [A], *TYR* [A]
MYC	*CHKA* [I], ***LDHA*** **[A]**, *ODC1* [A], *PRDX3* [A], *PRODH* [I]
NR1H4	*ABCB4* [A], *ABCC4* [I], *CYP7A1* [I], *FABP6* [A]
NRIP1	*SLC7A1* [A]
PAX6	*FABP7* [A], *PDHX* [I]
PPARA	*ACSL1* [I], *UCP1* [I]
PPARG	*ABCG2* [A], *ACAT1* [A], *CD36* [A], *FABP4* [A], *GK* [A], *SLC2A4* [I], *SLC5A5* [A], *SLC9A1* [I]
PPARGC1A	*ALDOB* [A], *CYP7A1* [A], *PCK2* [A]
SP1	*ABCA1* [A], *ABCC3* [A], *BSG* [A], *COX4I1* [I], *DHCR24* [A], *GCLC* [A], ***LDHA*** **[A]**, *MAT2A* [A], *MAT2B* [A], *NDUFV1* [A], *NDUFV2* [A], *P4HA1* [A], *PCK1* [A], *PHGDH* [A], *SLC19A1* [A], *SLC29A1* [I], *SLC5A1* [I], *SLC5A8* [A], *SLC9A3* [A], *SOD1* [A], *SOD1* [I], *SOD2* [I], TK1 [A], *UGDH* [A]
SP3	*ABCA1* [I], *BSG* [A], *SLC9A3* [A]
STAT1	*CFTR* [A], *CFTR* [I], *UPP1* [A]
STAT3	*NME1* [A], *UCP2* [I]
TP53	*CKM* [A], ***SLC2A1*** **[I]**, *SLC6A6* [I], *TYMS* [I]
ZNF143	*PCYT1A* [A]

Based on the discretized transcriptomics data from the breast cancer NCI60 cell lines, we associated a state (active or inactive) to the signaling species and to the metabolic genes that were part of the integrated model. In particular, the transcriptomics data had

information for 600 species in the upstream MYC signaling network, among them 468 were active species and 132 inactive species, and for the states of 376 metabolic genes that were part of the metabolic breast cancer model, specifically 242 active and 134 inactive (Table 4.3).

Next, we performed CONSIGN considering first only the signaling network, and then the integrated model containing both the signaling and the metabolic networks. We analyzed the differences in the results when we maximize the consistency between the network and the data imposing only the state defined by the transcriptomics data in the signaling species or including also the states of the metabolic genes.

When we applied CONSIGN only in the signaling network, we were able to consistently integrate the transcriptomics data to map the discretized expression of 531 proteins, and there existed four alternative solutions that simultaneously constrained the states for the maximum number of species consistently with the data. Analyzing the alternative solutions, we identified that the state of 529 proteins out of the 531 could always be consistent with the data (Table 4.3), in particular, 402 active and 127 inactive. The over-expressed or active proteins belong mainly to the following pathways: immune system, signal transduction, DNA repair, gene expression, cell cycle, and metabolism of proteins (Figure 4.4 A), which are commonly known to have higher activity in cancer cells. Small proportions of the 127 lowly-expressed or inactive species belong to several pathways, including metabolism, immune system, and hemostasis (Figure 4.4B).

We constrained these 529 proteins in the model to the corresponding state, and we used the model to analyze the states of other species in the network. To this end, we performed variability analysis in the states of all the signaling species. We identified 1625 additional species that require a tightly defined state to allow the model to meet the maximum consistency with the transcriptomics data. The constrained signaling model reveals that 644 of them should be active, and 981 of them inactive. Among the active species, the model correctly identified two phosphorylated species that were reported as active in the transcriptomics; 113 species that were not significantly deregulated in the data but the model shows evidence that they need to be active in order to represent the observed states of the other species; and 526 species that were not measured in the transcriptomics data. These results highlight the potential of the model to infer the state of additional species that were not part of the initial dataset but whose states are

indirectly defined by the other species. Finally, we identified three genes whose state in the discretized transcriptomics data was defined as inactive, but the model requires them as active, namely, MSH4, SGIP1, DMC1, which belong to the cell cycle and reproduction pathways.

Regarding the 981 inferred inactive species, the model correctly identified that phosphorylated PLA2G4A as inactive, 73 species that were not significantly down-regulated in the data, but they were characterized as inactive by the model, and 838 species that were not part of the transcriptomics data set. Moreover, we identified 69 species that are active in the transcriptomics data, but the model requires these species to be inactive.

We hypothesized three reasons that could explain the inconsistencies between the network states and the observed transcriptomics data: the threshold defined for the discretization of the states of the species based on the transcriptomics values, the incompleteness of the signaling network, or post-transcriptional and post-translational modifications. However, further analysis should be performed to validate and test these hypotheses.

Overall with our analysis, we could assign deregulation to the signaling subsystems related to the MYC upstream network. In particular, we observed activity related to the following pathways: immune system, signal transduction, metabolism, metabolism of RNA, some pathways in DNA repair, cell cycle, and cellular responses to external stimuli. Furthermore, we could assign a lower activity to a part of the signal transduction pathways, pathways related to DNA repair, and hemostasis.

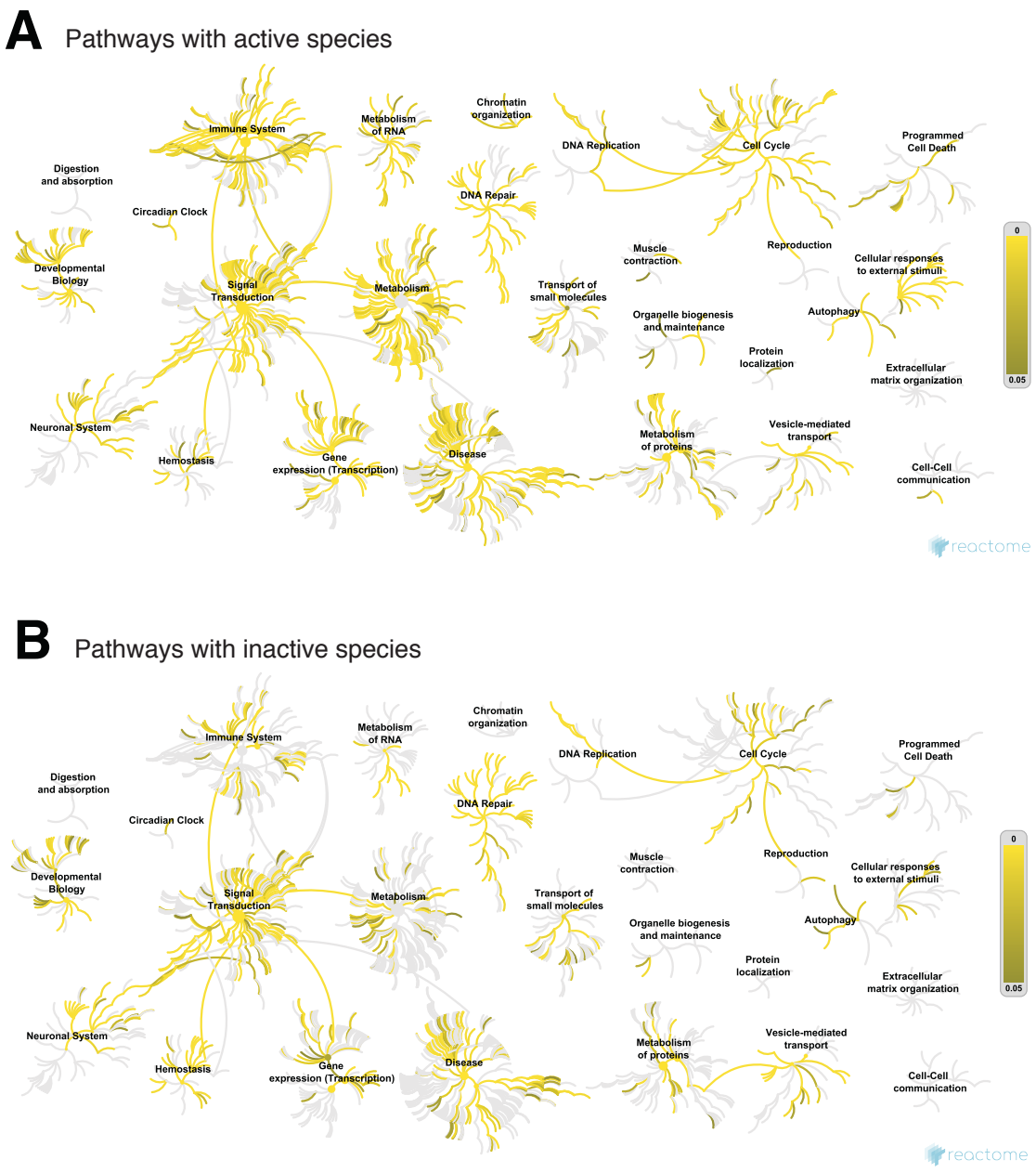


Figure 4.4. **Distribution of the network species that are constrained to be active or inactive based on the transcriptomics data and the signaling network constraints.** Genome-wide classification of the Reactome pathways related to active (A) and inactive (B) species whose states where integrated and inferred from the consistency between the transcriptomics data and the signaling network. The graph was obtained with Reactome Data Analysis Tool. The color code denotes the over-representation of the pathway based on the p-value. Light grey indicates pathways that are not significantly over-represented.

Next, we applied CONSIGN to the combined model containing the signaling model for the upstream MYC network and the metabolic breast cancer model. In this case, we could simultaneously integrate a maximum of 883 from the 976 states defined by the

transcriptomics data. We generated the existing 16 alternative solutions that maximize the consistency with the data, and we identified 879 proteins and genes whose state is consistent across the 16 alternatives. Specifically, we identified 394 active proteins, 126 inactive proteins, 231 active genes, and 128 inactive genes (Table 4.3).

Notice that the number of signaling proteins that could be consistently integrated decreased when we considered the integrated model containing the signaling and metabolic models. The proteins whose states could not be constrained simultaneously to satisfy the maximum number of transcriptomics data in the integrated model were SP1, CYP51A1, DHCR7, FASN, HDAC2, SUMO2,3-K386-TP53, ATF2, TP53, and PPARGC1A which are related to the immune system and the signal transduction pathways. The transcriptomics data identified PPARGC1A as inactive and the rest of them as active. However, our analysis suggests that the state of these proteins cannot be simultaneously constrained to satisfy the expression data for the other species.

Table 4.3. **Integration of data and consistency analysis**. Comparison of the maximum consistency and the number of alternatives

		MODELS		
		signaling	signaling and metabolism	metabolism
TRANSCRIPTOMICS DATA	proteins active	468	468	-
	proteins inactive	132	132	-
	genes active	-	242	242
	genes inactive	-	134	134
	TOTAL	**600**	**976**	**376**
CONSISTENCY WITH THE NETWORK	maximum consistency	531	883	361
	number of alternatives	4	16	2
	consistency across alternatives	529	879	360
	proteins active across alternatives	402	394	-
	proteins inactive across alternatives	127	126	-
	genes active across alternatives	-	231	231
	genes inactive across alternatives	-	128	129

Furthermore, performing the analysis considering only the metabolic network, the number of genes whose state was consistent with the experimental data was slightly higher than when we considered the integrated model (Table 4.3). This result, together with the lower consistency of the protein states in the integrated model, shows evidence that the two systems affect each other and that for a more realistic analysis, they should be studied as a whole instead of independently.

We next constrained the states of the 879 consistent species in the integrated model, and we performed variability analysis of the states of the other proteins. We identified 708 proteins and genes that must be active, and 1071 that must be inactive to allow the model to be consistent with the 879 species whose states where defined from the transcriptomics data. Among these 1779 species, the state of three proteins were correctly predicted, two active and one inactive; 232 were discretized as medium (neither active nor inactive) in the transcriptomics data, but the model requires 168 of them to be active and 64 of them to be inactive; additionally, 529 proteins were found to be active and 911 inactive, and these proteins were not measured in the transcriptomics data; finally, 9 species were discretized as inactive but the model assigned to them an active state and 75 species were active in the discretized transcriptomics but inactive in the model.

Overall, with our analysis, we could assign with higher confidence deregulation to the signaling and metabolic pathways. In particular, the integrated model revealed an upregulation of the immune system, metabolism, metabolism of RNA, gene expression, metabolism of proteins, disease associated pathways, cell cycle, DNA Replication and some pathways in DNA Repair among others. Furthermore, it showed a downregulation of signal transduction pathways, DNA repair, metabolism of proteins and some pathways from the cell cycle, among others.

Further analysis will be performed to investigate the differences between the deregulated pathways and assign functionality to these deregulations based on the interplay between the signaling and the metabolic network.

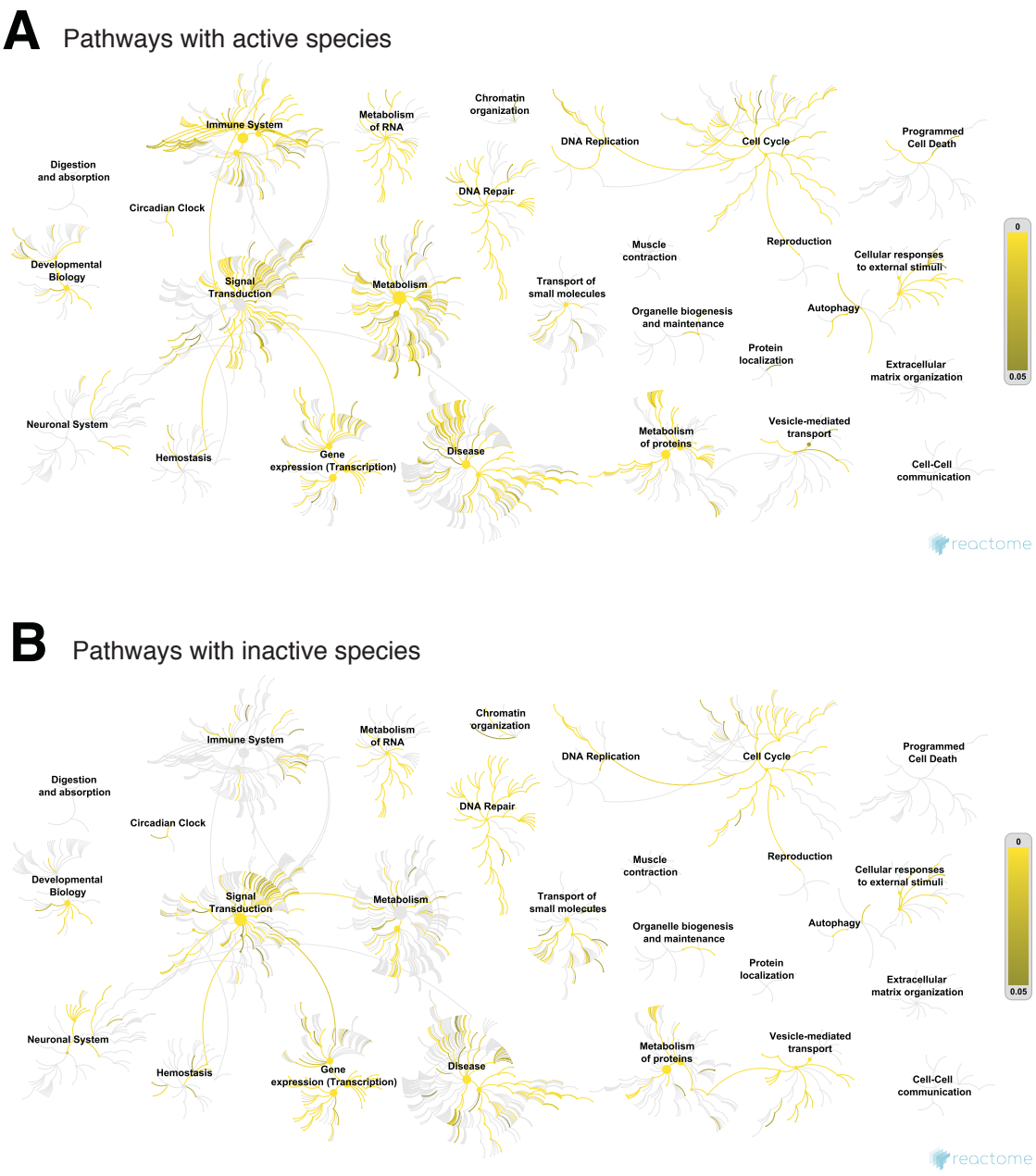


Figure 4.5. **Distribution of the network species that are constrained to be active or inactive based on the transcriptomics data and the integrated network constraints.** Genome-wide classification of the Reactome pathways related to active (A) and inactive (B) species whose states where integrated and inferred from the consistency between the transcriptomics data and the integrated network. The graph was obtained with Reactome Data Analysis Tool. The color code denotes the over-representation of the pathway based on the p-value. Light grey indicates pathways that are not significantly over-represented.

4.3 Discussion

A major challenge in systems biology is the integration of different biological networks and data-types. That is the case of signal transduction networks, gene expression, and metabolic networks, which have been mainly studied independently despite being closely related as they tightly regulate each other. We propose a new method named CONSIGN to contextualize signaling networks to a specific physiology. The method maximizes the consistency of the state of the species in the network and the experimental data. Furthermore, we present a workflow to generate integrated models of signaling and metabolic networks. The work here presented gives the opportunity to simultaneously simulate the state of the signaling and the metabolic networks and to integrate omics data, including transcriptomics, proteomics, metabolomics, and fluxomics, ensuring the consistency with both biological networks and with the experimental measurements. The integrated model is a powerful tool to analyze how the regulatory effects propagate to metabolism and to study critical points in the signal transduction pathway that can be investigated as targets for therapies. In particular, we can identify possible drug targets and the impact that the specific drugs would have in the network, accounting for alternative pathways that cells have to overcome the effect of the drugs generating resistance to the treatment.

In this work, we modeled the signaling network as a Boolean network, considering only two possible states for the signaling species, active or inactive. Although this is a strong assumption that is far from reality, the generated integrated model can be used for exploratory analysis to define the structure of the model. Moreover, it serves as a basis to develop an improved model, including the formulation of continuous or multilevel constraints to account for intermediate states of the signaling species and genes, and the refinement of the interactions between the metabolic and the signaling pathways [23, 58, 59]. Furthermore, in this analysis we have only considered the effect of the signaling pathways in the metabolic network. However, in the future, the integrated network could also take into account the regulatory effects of metabolism in the signaling pathways.

Herein, we focused on the transcription factor MYC and its effects on metabolism. By building the upstream signaling network related to MYC, we closely represented its regulatory network, including the impact of other signal transduction pathways. In further studies, we could include other signaling species that are regulated by MYC and capture

the propagation of the signal in the downstream signaling pathways from MYC, possibly including other metabolic regulators that could be connected to the metabolic network. For example, the repression by MYC of the microRNAs miR-23a/b that increase the expression of the glutaminase protein [60] affecting glutamine metabolism, which is indeed another of the alterations reported for cancer cells [61, 62].

As a case study and for consistency with the metabolic model generated in chapter 2, we used in this analysis the transcriptomics data from the NCI60 cell lines. Nevertheless, the same workflow can be applied integrating omics data from other datasets such as the Cancer Cell Line Encyclopedia (CCLE) [63], or The Cancer Genome Atlas Program (TCGA) [64] as well as patient-specific *omics* data. Phosphoproteomics and proteomics data can also be directly integrated into the generated model to account for the activity of the signaling proteins and enzymes with higher certitude. Moreover, to illustrate the applicability of the method and workflow, we have created a breast cancer integrated model. However, the same workflow can be applied to generate specific models per cell line or for different types or stages of breast cancer. Finally, the workflow applies to any other disease as well as to healthy cells having the opportunity to compare the signaling and metabolic differences and similarities between the diseased cells and their healthy counterparts.

In the case of using the workflow with patient-specific data, we have the opportunity to perform biomarker analysis by identifying the consistent states among the samples from the patients, the integrated network, and the alternative states of the network.

Overall, our study shows promising applications in the of systems biology and medicine bridging the studies of signaling and metabolism and allowing to refine the hypotheses suggested by previous studies and to test and design new experimental hypotheses.

4.4 Materials and Methods

4.4.1 Gene and protein expression data

Transcriptomics data were collected from the NCI60 cell lines for breast cancer. The transcriptomics data levels were discretized in three levels: high medium and low, based on the population of the expression levels for each species across cell lines. For each cell line, each species was classified as highly or lowly expressed if their expression levels were higher than the third quartile of the population or smaller than the first quartile of the population, respectively. For the specific cancer type, we assigned the corresponding expression to the genes if their deregulation was consistent in at least 80% of the samples for the cell lines of breast cancer.

The discretized transcriptomics data were used to determine the expression level of the signaling proteins and genes and the metabolic genes. In this work, we considered that highly expressed proteins and genes were active proteins, and lowly expressed proteins and genes were inactive proteins.

4.4.2 Human breast cancer metabolic model

The metabolic model is a breast-cancer specific reduced model of the thermodynamically curated human genome-scale model Recon 3D [41], reconstructed following the redHUMAN pipeline described in chapter 2. The reduced model was generated around 11 subsystems that have been reported to be altered in cancer, namely, 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. The reduced model accounts for the pathways required to uptake and secrete the extracellular metabolites based on the *omics* data and for the necessary pathways to synthesize the biomass building blocks, such as amino acids, nucleic acids, lipids, and proteins. Furthermore, we defined the physiology for breast cancer by integrated metabolomics and transcriptomics data from the NCI60 breast cancer cell lines [50] into the metabolic reduced model following the workflow described in chapter 3. The breast-cancer specific metabolic model consists of 613 metabolites, 1717 reactions, and 971 genes associated with the reactions.

4.4.3 Mapping transcription factors to metabolism

We obtained information about transcription factors (TFs) and their target genes from the TRRUST database [57] (https://www.grnpedia.org/trrust/). In particular, we identified a total of 5071 pairs composed of 795 transcription factors and 2492 genes and their corresponding regulation, that is, activation or repression. Next, we identified the genes from the metabolic network Recon 3D [41], whose expression was regulated by a transcription factor. As a result, we identified 460 pairs of 115 transcription factors and 237 metabolic genes from Recon 3D and the corresponding regulatory effect. From those 460 pairs, 180 pairs formed of 77 transcription factors that regulate a total of 117 genes are associated with the breast cancer-specific reduced model used in this work (Table S1).

4.4.4 Reconstruction of signaling pathways from REACTOME

We extracted from Reactome [3] (https://reactome.org/) the information of the signaling pathways, including the reactions, the species, and the regulatory effects such as activation and repression of the signaling reactions.

We built a directed graph by identifying the species and reactions as nodes and the interactions as directed edges. As a result, substrates/modifiers and regulators are connected through an edge to their corresponding reactions, and reactions are connected through an edge to their corresponding products. The directed graph is then navigated to obtain signaling pathways for specific nodes. For a signaling species, we perform a directed graph search to find all the predecessors in the directed graph and build their upstream network or to find all the successors resulting in the downstream network. For the graph search, we did not take into account cofactors, to maintain only the main signaling pathway.

In this work, we built the MYC signaling pathway by identifying all its predecessors and all the predecessors of each predecessor in the directed graph. The final graph has 117 layers of predecessors.

4.4.5 Discrete formulation of rules for GPRs

The gene-protein-reaction (GPRs) rules in the models describe the relationship among genes, enzymes, and the activity of the reactions they catalyze using Boolean logic rules

with the standard operators AND and OR. We used the TIGER toolbox [56] to convert the GPR Boolean rules into algebraic equations that can be integrated into the optimization problem (MILP). In this context, the following transformations for simple rules are applied:

$$x \Leftrightarrow z \quad \longrightarrow \quad x - z = 0 \tag{1}$$

$$x \text{ AND } y \Leftrightarrow z \quad \longrightarrow \quad \begin{cases} 2x + 2y - 4z \geq -1 \\ 2x + 2y - 4z \leq 3 \end{cases} \tag{2}$$

$$x \text{ OR } y \Leftrightarrow z \quad \longrightarrow \quad \begin{cases} -x - y + 3z \geq 0 \\ -x - y + 3z \leq 2 \end{cases} \tag{3}$$

where x and y are binary variables representing the state of genes, and z is a binary variable that represents the activity of the reaction. Thus, if the GPR is true, then $z = 1$ and the reaction can carry flux.

In the case of more complex rules, auxiliary variables I are introduced to formulate the inequalities (see [56] for further details), for example:

$$(x \text{ AND } y) \text{ OR } s \Leftrightarrow z \quad \longrightarrow \quad \begin{cases} 2x + 2y - 4I \leq 3 \\ 2x + 2y - 4I \geq -1 \\ -I - s + 3z \geq 0 \\ -I - s + 3z \leq 2 \end{cases} \tag{4}$$

where x, y, z, I, and s are binary variables.

We add the linear inequalities as constraints to the MILP formulation of the TFA problem for the metabolic model. In addition, we add the following constraints to link the state of the reaction to the corresponding net flux.

$$z - b^{NF} = 0 \tag{5}$$

$$v^{NF} - C \cdot b^{NF} \leq 0 \tag{6}$$

$$v^{NF} + C \cdot b^{NF} \geq 0 \tag{7}$$

where z is the binary variable representing the activity of the reaction and b^{NF} is a new binary variable that controls the net flux through the reaction in the TFA problem. Moreover, v^{NF} is the TFA net flux variable for the reaction, related to the forward and backward fluxes through the TFA formulation [51, 65] and C an arbitrary big number (larger than the highest flux in the network). In this work, $C = 10^6$. With these constraints,

if $z = 1$ then $b^{NF} = 1$ and $v^{NF} \in (-C, C)$. On the opposite, if based on the GPR the reaction is inactive then $b^{NF} = 0$ and $v^{NF} = 0$ blocking the flux through the corresponding reaction.

4.4.6 Discrete formulation of rules for signaling interactions

We formulated Boolean rules to describe the signaling interactions including activation, repression, complex formation and product formation. We consider as product the modified species by the reaction for example the (de)phosphorylated or (de)ubiquitinated protein. To this end, we formulated rules that activate the reaction in the presence of substrate and activators and in the absence of repressors and rules that activate the products if the reaction is active. Specifically, we used the AND operator and the not operator to describe repression. A general set of rules were formulated as it follows:

$$substrate \text{ AND } activator \text{ AND not } repressor \rightarrow reaction$$
$$reaction \rightarrow product$$

Then, we converted the Boolean rules into algebraic equations that can be integrated as constraints in the optimization problem. In this context, for the previous general rule, the following transformations were applied:

$$\begin{gathered} r + r^{not} = 1 \\ 2s + 2a - 4I \leq 3 \\ 2s + 2a - 4I \geq -1 \\ 2I + 2r^{not} - 4R \geq 3 \\ 2I + 2r^{not} - 4R \leq -1 \\ R - p = 0 \end{gathered} \tag{8}$$

Where r, s, a, R, and p are binary variables that represent the state (on or off) of the repressor, the substrate, the activator, the reaction, and the product, respectively. r^{not}, and I are auxiliary binary variables to formulate the not operator and the rule.

We formulate a mixed-integer linear programming optimization problem with binary variables representing the states of the species and reactions, and the rules as constraints.

4.4.7 Discrete formulation of rules for the regulation of genes by transcription factors

To captured how transcription factors regulate the expression of genes, we formulated Boolean rules that can be summarized in the following cases:

1. The expression of a gene is promoted by one transcription factor: $TF \Leftrightarrow gene$.
2. The expression of a gene is repressed by one transcription factor: $\text{not}\, TF \Leftrightarrow gene$.
3. The expression of a gene can be promoted by several transcription factors: $TF_1 \,\text{OR}\, TF_2 \Leftrightarrow gene$.
4. The expression of a gene is promoted by a transcription factor and inhibited by another transcription factor: $TF_1 \,\text{AND not}\, TF_2 \Leftrightarrow gene$.

We then converted the Boolean rules into algebraic equations as it follows: the first case was modeled using equation (1); the second case was modeled using equation (1) and the auxiliary rule for the not operator, $x + x^{not} = 1$. The third case was modeled using equation (3) and the fourth case was modeled using equation (2) and the auxiliary rule for the not operator.

The algebraic equations are added to the MILP together with the signaling and the metabolic models. These rules bridge the regulatory effects between the signaling networks and the metabolic networks.

4.4.8 Contextualization of signaling networks (CONSIGN) method

CONSIGN is a Boolean method that generates context-specific signaling networks and identifies signaling components (i.e. proteins and genes) whose activity in the network is consistent with experimental data. The method can as well identify active or inactive signaling reactions based on the activity of its participants. CONSIGN translates a given signaling network into a set of Boolean rules and further into a mixed integer programming problem (MILP) using the discrete formulation of rules from the TIGER toolbox.

Moreover, the algorithm allows us to integrate transcriptomics or proteomics data of a given context and identify the signaling components and events that are maximally consistent with the context-specific data. To this end, we formulated the following

optimization problem (MILP), where the objective is to maximize consistency between the measured target states and the network:

$$max \sum_{k \in (S_A \cup S_R)} b_k$$

s.t.

$$Equations\ for\ Signaling\ Rules$$

$$x_i - b_i > 0, \quad \forall i \in S_A$$

$$y_j + 100 \cdot b_j < 100, \quad \forall j \in S_R$$

where S_A and S_R represent the set of active and repressed species based on the data, x and y are the binary variables introduced with the equations that represent the signaling rules, b are new binary variables representing the state of the species. For the species that are active based on the data, if $b_i = 1$ then the corresponding species should be active in the model, while for the species that are inactive according to the experimental data, if $b_j = 1$ then the constraint forces the corresponding species in the model to be inactive. When considering the combined signaling and metabolic models, we include a constraint to set the biomass to its maximum value, $v_{biomass}^{LB} = v_{biomass,max}$.

The optimization problem maximizes the number of $b_k = 1$, maximizing the consistency between the simulated states of the network and the experimental data. The MILP formulation allows generating alternatives, enabling to explore how the network is able to accommodate the state of the species according to the data.

Finally, we define in the network the states of the species that are always consistent across alternatives, and we study the states of the network that can explain the data. To this end, we formulate the following MILP to maximize the active states of the network.

$$max \sum_{s \in S} b_s$$

s.t.

$$Equations\ for\ Signaling\ Rules$$

$$x_i = 1, \quad \forall i \in S_{C,A}$$

$$y_j = 0, \quad \forall j \in S_{C,R}$$

where S is the set of species in the network, $S_{C,A}$ and $S_{C,R}$ are the set of active and inactive species, respectively, according to the data and whose state is consistent with the signaling network. We included additional constraints in the MILP formulation to enumerate all the possible alternative states of the network that are consistent with the integrated data.

4.5 Author contribution

The work from this chapter is *in preparation* to be published with the provisory title: *Integrating signaling and metabolic pathways to analyze the function of the transcription factor MYC in breast cancer.*

For this work, the literature collection and processing of the data, the generation of the metabolic models, the generation of the signaling models, the generation of the integrated metabolic-signaling models, the writing of the current manuscript and chapter, as well as, all the figures were performed by Maria Masid under the supervision of Prof. Vassily Hatzimanikatis. The conceptualization of the method CONSIGN as well as the software/code was initially developed by Dr. Vikash Pandey and further improved by Maria Masid, both under the supervision of Prof. Vassily Hatzimanikatis. The extraction and visualization of signaling networks from Reactome was designed by Vikash Pandey, and improved and performed by Maria Masid assisted by Evangelia Vayena, all under the supervision of Prof. Vassily Hatzimanikatis.

References

1. Alberts, B., et al., *Molecular Biology of the Cell*. 4th ed. 2002: Garland Science.
2. Jordan, J.D., E.M. Landau, and R. Iyengar, *Signaling networks: the origins of cellular multitasking.* Cell, 2000. **103**(2): p. 193-200.
3. Fabregat, A., et al., *The Reactome pathway Knowledgebase.* Nucleic Acids Research, 2016. **44**(D1): p. D481-D487.
4. Kanehisa, M., et al., *KEGG: new perspectives on genomes, pathways, diseases and drugs.* Nucleic Acids Research, 2017. **45**(D1): p. D353-D361.
5. Schaefer, C.F., et al., *PID: the Pathway Interaction Database.* Nucleic Acids Research, 2009. **37**: p. D674-D679.
6. Kandasamy, K., et al., *NetPath: a public resource of curated signal transduction pathways.* Genome Biology, 2010. **11**(1).
7. Kuperstein, I., et al., *Atlas of Cancer Signalling Network: a systems biology resource for integrative analysis of cancer data with Google Maps.* Oncogenesis, 2015. **4**.
8. Janes, K.A. and D.A. Lauffenburger, *Models of signalling networks - what cell biologists can gain from them and give to them.* Journal of Cell Science, 2013. **126**(9): p. 1913-1921.
9. Chen, R.E. and J. Thorner, *Systems biology approaches in cell signaling research.* Genome Biology, 2005. **6**(10).
10. Nair, A., et al., *Conceptual Evolution of Cell Signaling.* International Journal of Molecular Sciences, 2019. **20**(13).
11. Sebastian-Leon, P., et al., *Understanding disease mechanisms with models of signaling pathway activities.* Bmc Systems Biology, 2014. **8**.
12. Kondratova, M., et al., *A multiscale signalling network map of innate immune response in cancer reveals cell heterogeneity signatures.* Nature Communications, 2019. **10**.
13. Choi, S., *Systems biology for signaling networks*. Systems biology. 2010, New York: Springer. xvi, 908 p.

14. Karlebach, G. and R. Shamir, *Modelling and analysis of gene regulatory networks.* Nature Reviews Molecular Cell Biology, 2008. **9**(10): p. 770-780.

15. De Jong, H., *Modeling and simulation of genetic regulatory systems: A literature review.* Journal of Computational Biology, 2002. **9**(1): p. 67-103.

16. Le Novere, N., *Quantitative and logic modelling of molecular and gene networks.* Nature Reviews Genetics, 2015. **16**(3): p. 146-158.

17. Neves, S.R. and R. Iyengar, *Modeling of signaling networks.* Bioessays, 2002. **24**(12): p. 1110-7.

18. Hyduke, D.R. and B.O. Palsson, *Towards genome-scale signalling-network reconstructions.* Nature Reviews Genetics, 2010. **11**(4): p. 297-307.

19. Wang, R.-S., A. Saadatpour, and R. Albert, *Boolean modeling in systems biology: an overview of methodology and applications.* Physical Biology, 2012. **9**: p. 055001.

20. Samaga, R., et al., *The Logic of EGFR/ErbB Signaling: Theoretical Properties and Analysis of High-Throughput Data.* Plos Computational Biology, 2009. **5**(8).

21. Martin, S., et al., *Boolean dynamics of genetic regulatory networks inferred from microarray time series data.* Bioinformatics, 2007. **23**(7): p. 866-874.

22. Barman, S. and Y.K. Kwon, *A Boolean network inference from time-series gene expression data using a genetic algorithm.* Bioinformatics, 2018. **34**(17): p. 927-933.

23. Albert, R. and J. Thakar, *Boolean modeling: a logic-based dynamic approach for understanding signaling and regulatory networks and for making useful predictions.* Wiley Interdisciplinary Reviews: Systems Biology and Medicine, 2014. **6**(5): p. 353-69.

24. Sharan, R. and R.M. Karp, *Reconstructing Boolean Models of Signaling.* Journal of Computational Biology, 2013. **20**(3): p. 249-257.

25. Razzaq, M., et al., *Computational discovery of dynamic cell line specific Boolean networks from multiplex time-course data.* PLoS Comput Biol, 2018. **14**(10): p. e1006538.

26. Saez-Rodriguez, J., et al., *Discrete logic modelling as a means to link protein signalling networks with functional analysis of mammalian signal transduction.* Molecular Systems Biology, 2009. **5**: p. 331.

27. Singh, A., et al., *Boolean approach to signalling pathway modelling in HGF-induced keratinocyte migration.* Bioinformatics, 2012. **28**(18): p. I495-I501.

28. Leifeld, T., Z. Zhang, and P. Zhang, *Identification of Boolean Network Models From Time Series Data Incorporating Prior Knowledge.* Frontiers in Physiology, 2018. **9**: p. 695.

29. Fumia, H.F. and M.L. Martins, *Boolean Network Model for Cancer Pathways: Predicting Carcinogenesis and Targeted Therapy Outcomes.* Plos One, 2013. **8**(7).

30. Sridharan, S., et al., *Hypoxia Stress Response Pathways: Modeling and Targeted Therapy.* Ieee Journal of Biomedical and Health Informatics, 2017. **21**(3): p. 875-885.

31. Saez-Rodriguez, J. and N. Bluthgen, *Personalized signaling models for personalized treatments.* Mol Syst Biol, 2020. **16**(1): p. e9042.

32. Zanudo, J.G.T., S.N. Steinway, and R. Albert, *Discrete dynamic network modeling of oncogenic signaling: Mechanistic insights for personalized treatment of cancer.* Current Opinion in Systems Biology, 2018(9): p. 10-11.

33. Saez-Rodriguez, J., A. MacNamara, and S. Cook, *Modeling Signaling Networks to Advance New Cancer Therapies.* Annu Rev Biomed Eng, 2015. **17**: p. 143-63.

34. Zhang, R.R., et al., *Network model of survival signaling in large granular lymphocyte leukemia.* Proceedings of the National Academy of Sciences of the United States of America, 2008. **105**(42): p. 16308-16313.

35. Zhang, F., R.S. Liu, and J. Zheng, *Sig2GRN: a software tool linking signaling pathway with gene regulatory network for dynamic simulation.* Bmc Systems Biology, 2016. **10**.

36. Oster, S.K., et al., *The myc oncogene: MarvelousIY Complex.* Adv Cancer Res, 2002. **84**: p. 81-154.

37. Dang, C.V., *MYC on the Path to Cancer.* Cell, 2012. **149**(1): p. 22-35.

38. Soucek, L. and N.M. Sodir, *The Myc gene : methods and protocols*. Methods in molecular biology,. 2013, New York: Humana Press ; Springer. xi, 282 p.

39. McKeown, M.R. and J.E. Bradner, *Therapeutic strategies to inhibit MYC.* Cold Spring Harb Perspect Med, 2014. **4**(10).

40. Dang, C.V., A. Le, and P. Gao, *MYC-induced cancer cell energy metabolism and therapeutic opportunities.* Clin Cancer Res, 2009. **15**(21): p. 6479-83.

41. Brunk, E., et al., *Recon3D enables a three-dimensional view of gene variation in human metabolism.* Nature Biotechnology, 2018. **36**(3): p. 272-+.

42. Robinson, J.L., et al., *An atlas of human metabolism.* Science Signaling, 2020. **13**(624).

43. Covert, M.W., C.H. Schilling, and B. Palsson, *Regulation of gene expression in flux balance models of metabolism.* Journal of Theoretical Biology, 2001. **213**(1): p. 73-88.

44. Shlomi, T., et al., *A genome-scale computational study of the interplay between transcriptional regulation and metabolism.* Molecular Systems Biology, 2007. **3**.

45. Covert, M.W., et al., *Integrating metabolic, transcriptional regulatory and signal transduction models in Escherichia coli.* Bioinformatics, 2008. **24**(18): p. 2044-2050.

46. Chandrasekaran, S. and N.D. Price, *Probabilistic integrative modeling of genome-scale metabolic and regulatory networks in Escherichia coli and Mycobacterium tuberculosis.* Proceedings of the National Academy of Sciences of the United States of America, 2010. **107**(41): p. 17845-17850.

47. Wang, Z., et al., *Combining inferred regulatory and reconstructed metabolic networks enhances phenotype prediction in yeast.* PLoS Comput Biol, 2017. **13**(5): p. e1005489.

48. Marmiesse, L., R. Peyraud, and L. Cottret, *FlexFlux: combining metabolic flux and regulatory network analyses.* Bmc Systems Biology, 2015. **9**.

49. Motamedian, E., et al., *TRFBA: an algorithm to integrate genome-scale metabolic and transcriptional regulatory networks with incorporation of expression data.* Bioinformatics, 2017. **33**(7): p. 1057-1063.

50. Jain, M., et al., *Metabolite Profiling Identifies a Key Role for Glycine in Rapid Cancer Cell Proliferation.* Science, 2012. **336**(6084): p. 1040-1044.

51. Henry, C.S., L.J. Broadbelt, and V. Hatzimanikatis, *Thermodynamics-based metabolic flux analysis.* Biophysical Journal, 2007. **92**(5): p. 1792-1805.

52. Evan, G.I. and T.D. Littlewood, *The role of c-myc in cell growth.* Curr Opin Genet Dev, 1993. **3**(1): p. 44-9.

53. Hoffman, B. and D.A. Liebermann, *Apoptotic signaling by c-MYC.* Oncogene, 2008. **27**(50): p. 6462-72.

54. Adhikary, S. and M. Eilers, *Transcriptional regulation and transformation by Myc proteins.* Nat Rev Mol Cell Biol, 2005. **6**(8): p. 635-45.

55. Miller, D.M., et al., *c-Myc and cancer metabolism.* Clin Cancer Res, 2012. **18**(20): p. 5546-53.

56. Jensen, P.A., K.A. Lutz, and J.A. Papin, *TIGER: Toolbox for integrating genome-scale metabolic models, expression data, and transcriptional regulatory networks.* BMC Syst Biol, 2011. **5**: p. 147.

57. Han, H., et al., *TRRUST v2: an expanded reference database of human and mouse transcriptional regulatory interactions.* Nucleic Acids Res, 2018. **46**(D1): p. D380-D386.

58. Morris, M.K., et al., *Logic-Based Models for the Analysis of Cell Signaling Networks.* Biochemistry, 2010. **49**(15): p. 3216-3224.

59. Klamt, S., et al., *A methodology for the structural and functional analysis of signaling and regulatory networks.* BMC Bioinformatics, 2006. **7**: p. 56.

60. Gao, P., et al., *c-Myc suppression of miR-23a/b enhances mitochondrial glutaminase expression and glutamine metabolism.* Nature, 2009. **458**(7239): p. 762-U100.

61. Wise, D.R. and C.B. Thompson, *Glutamine addiction: a new therapeutic target in cancer.* Trends in Biochemical Sciences, 2010. **35**(8): p. 427-433.

62. Altman, B.J., Z.E. Stine, and C.V. Dang, *From Krebs to clinic: glutamine metabolism to cancer therapy.* Nat Rev Cancer, 2016. **16**(11): p. 749.

63. Barretina, J., et al., *The Cancer Cell Line Encyclopedia enables predictive modelling of anticancer drug sensitivity.* Nature, 2012. **483**(7391): p. 603-607.

64. Tomczak, K., P. Czerwinska, and M. Wiznerowicz, *The Cancer Genome Atlas (TCGA): an immeasurable source of knowledge.* Contemp Oncol (Pozn), 2015. **19**(1A): p. A68-77.

65. Soh, K.C.a.H., V., *Constraining the flux space using thermodynamics and integration of metabolomics data.* Methods Mol Biol (Clifton, N.J.), 2014(1191): p. pp. 49-63.

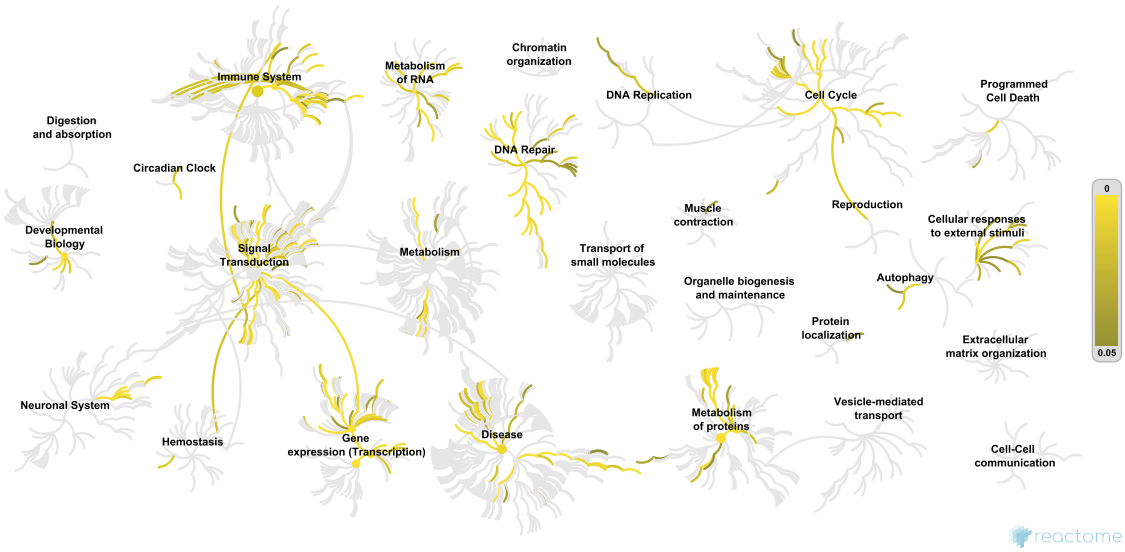
Appendix C

Table S1: **Regulation of metabolic genes by transcription factors**. List of transcription factors and the genes that they regulate and that are part of the breast cancer specific reduced metabolic model.

TFs	Genes [Regulation]
AHR	CA9 [I]
APBB1	GMPS [I]
APC	ODC1 [I]
APEX1	SLC5A5 [A]
AR	ARG1 [A], ARG2 [A], PC [A]
ARNT	CA9 [A]
ATF1	CFTR [A], SLC20A1 [A]
ATF2	PCK1 [A]
ATF3	ASNS [A]
ATM	SLC2A1 [I]
CDCA7L	MAOB [I]
CDX2	ACAT2 [A], CFTR [A], SLC15A1 [A]
CEBPB	GOT1 [A], PCK2 [A], SLC19A1 [A], SLC5A8 [A]
CEBPD	SOD1 [A]
CREB1	ACACA [A], SLC20A1 [A]
CREB3L1	SLC1A4 [I]
CREB5	DGKG [A], DGKG [I]
CREM	G6PD [I], SLC5A5 [A], SLC5A5 [I]
CTNNB1	PLD1 [A]
DBP	ALDOB [A], CYP7A1 [A]
DDB2	SOD2 [I]
E2F1	DHFR [A], ISYNA1 [A], RRM1 [A], RRM2 [A], TYMS [A]
EGR1	SLC4A2 [A], SLC9A3 [A], SOD1 [A]
EGR3	CAT [A]
EZH2	ALDH1A1 [I]
FOS	SLC10A2 [I]
GATA1	CDA [A]
GATA2	ADH1A [A]
GATA4	FABP2 [A]
HDAC1	GAD1 [I]
HDAC2	PRDX2 [I]
HDAC7	HDC [I]
HIF1A	ACE2 [I], ALDOA [A], CA9 [A], LDHA [A], NT5E [A], PGK1 [A], SDHB [I], SLC29A1 [I], SLC2A1 [A]
HMGA1	SLC2A3 [A]
HNF1A	ACAT2 [A]
HNF4A	ABCG5 [I], ABCG8 [I], CYP7A1 [I], FABP2 [A]
HOXA10	PHGDH [I]
IRF3	ABCC2 [A]
JUN	LDHA [A], MAT2A [A]
KLF2	FABP5 [I]

KLF3	CKM [A]
KLF5	FASN [A]
MEF2A	SLC2A4 [A]
MITF	ACP5 [A], TYR [A]
MTF1	GCLC [A], SOD1 [A]
MYB	CDO1 [A], MAT2A [A], TK1 [A]
MYC	CHKA [I], LDHA [A], ODC1 [A], PRDX3 [A], PRODH [I]
MYCN	ABCC1 [A]
NFATC1	CYP2E1 [A]
NFE2L2	CAT [A], CFTR [I], GCLC [I], MTHFR [A], SOD1 [A], SOD2 [A]
NFIC	SLC34A2 [A]
NFKB1	ABCA1 [I], ABCG2 [A], CFTR [A], MAT2A [A], SLC25A27 [A], SOD2 [A], UPP1 [A]
NFYA	CBS [A]
NR0B2	PCK2 [I]
NR1H4	ABCB4 [A], ABCC4 [I], CYP7A1 [I], FABP6 [A]
NR3C1	ATP1B1 [I]
NRF1	PRDX3 [A], SLC46A1 [A]
NRIP1	SLC7A1 [A]
PAX6	FABP7 [A], PDHX [I]
PITX3	MIP [A]
POU2F1	TK1 [A]
PPARA	ACSL1 [I], UCP1 [I]
PPARG	ABCG2 [A], ACAT1 [A], CD36 [A], FABP4 [A], GK [A], SLC2A4 [I], SLC5A5 [A], SLC9A1 [I]
PTMA	IDO1 [A], IDO2 [A]
RARA	ABCC3 [I], SCD [A]
SP1	ABCA1 [A], ABCC3 [A], BSG [A], COX4I1 [I], DHCR24 [A], GCLC [A], LDHA [A], MAT2A [A], MAT2B [A], NDUFV1 [A], NDUFV2 [A], P4HA1 [A], PCK1 [A], PHGDH [A], SLC19A1 [A], SLC29A1 [I], SLC5A1 [I], SLC5A8 [A], SLC9A3 [A], SOD1 [A], SOD1 [I], SOD2 [I], TK1 [A], UGDH [A]
SP3	ABCA1 [I], BSG [A], SLC9A3 [A]
SREBF1	ACLY [A], PCK1 [I]
SREBF2	ABCA1 [I], ABCG5 [I], ABCG8 [I], HMGCR [A]
STAT3	NME1 [A], UCP2 [I]
TFAP2A	NME3 [A], SLC19A1 [A]
TFAP2C	GPX1 [A]
TP53	CKM [A], SLC2A1 [I], SLC6A6 [I], TYMS [I]
TWIST1	CTPS1 [A]
USF1	ABCA1 [A], GCK [A], SLC19A1 [A], SLC22A2 [A]
YY1	CFTR [A], COX7C [A]
ZNF143	PCYT1A [A]

A Pathways with consistently active species for the upstream MYC signaling network



B Pathways with consistently inactive species for the upstream MYC signaling network

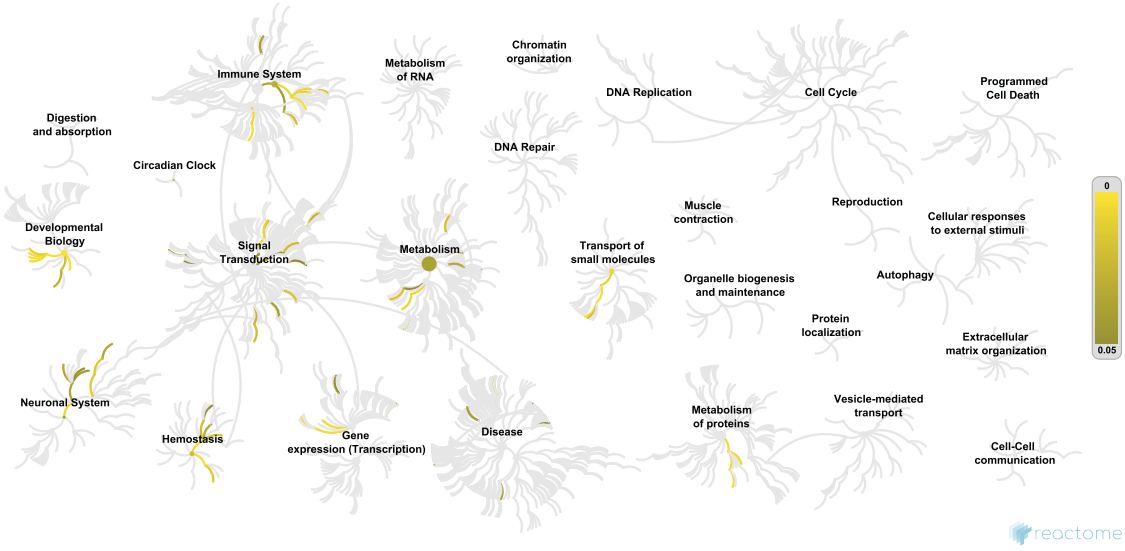
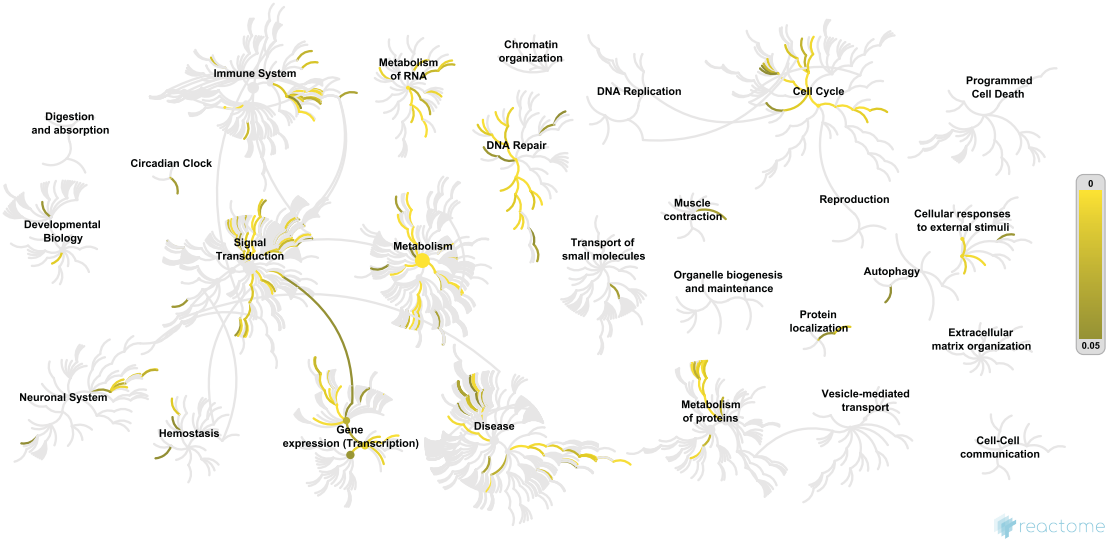


Figure S4.1. **Overall distribution of the state of the species that maximize consistency between the signaling network and the data.**

A Pathways with consistently active species for the integrated model



B Pathways with consistently inactive species for the integrated model

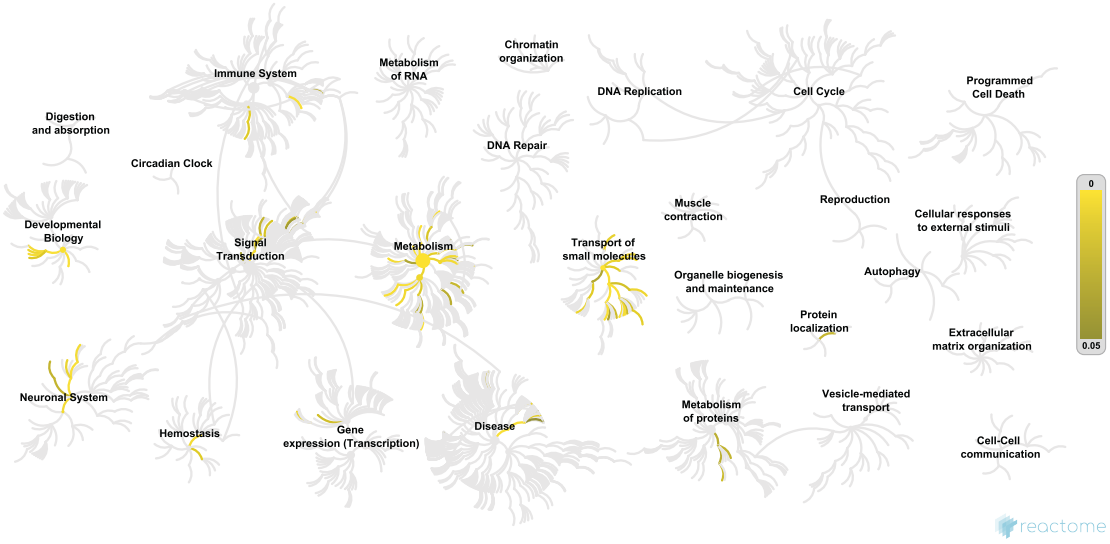


Figure S4.2. **Overall distribution of the state of the species that maximize consistency between the integrated network and the data.**

Chapter 5 Conclusions

In this chapter, we summarize the work and the main findings presented in the thesis, and we discuss their relevance and their contributions to the field of Systems Biology as well as future perspectives of applications of the models and methods here developed.

5.1 Conclusions

The study of cancer as a systems disease, where we do not focus on individual parts but in the overall behavior within and across cellular processes, is crucial to have a deeper understanding of the underlying alterations that define the observed phenotype. Systems biology approaches and techniques are needed to develop improved models and methods that allow us to investigate this disease and to propose new targets for therapies. The work in this thesis contributes to overcoming some of the current challenges in systems biology that hinder the analysis of biological networks in complex systems as human cells.

In **Chapter 2**, we overcame the challenge of employing large, complex metabolic models to study the metabolic physiology of human cells. We performed a thermodynamic curation of the genome-scale models for human metabolism, Recon 2 [1] and Recon 3D [2] including the thermodynamic properties for the compounds and reactions ensuring that the directionality of the reactions and the concentrations of the metabolites are consistent with the bioenergetics of the cell and with the laws of thermodynamics [3, 4]. We analyzed and classified the extracellular medium required to simulate growth, allowing to characterize the compounds assigned to the extracellular space in the model.

This analysis allows to define a biologically relevant medium in the model and refine the predictions as we constrain the set of metabolites that the simulated cell can uptake. Next, we developed a workflow (redHUMAN) that based on the methods redGEM [5] and lumpGEM [6] allows to reduce the size, and thus the complexity, of the human GEMs, and generate reduced models that focus on specific parts of the metabolism that are relevant for the physiology under study. The models created with redHUMAN include, in addition to the initial set of pathways, the metabolic pathways that the cells use to interact with a defined extracellular medium, and the metabolic pathways that the cells need to biosynthesize the cellular building blocks. The addition of these two sets of pathways allows to simulate with the reduced model a complete physiological scenario. Finally, the models undergo a list of consistency checks that include the capability of performing a set of metabolic tasks defined for human cells, ensuring that the predictive capabilities of the GEM, in terms of growth, reaction directionality, and gene essentiality are conserved in the generated reduced model. The reduced models are powerful platforms for studying metabolic differences between phenotypes, such as diseased and healthy cells.

In **Chapter 3**, we demonstrate the relevance of the generated reduced models to investigate the metabolism of cancer cells and to translate the observed deregulation at the genetic level into deregulation of the metabolic pathways related to the genes. We integrated exo-metabolomics, exo-fluxomics, transcriptomics data from the NCI60 cancer cell lines [7] into the generated reduced model from Recon 3D to build metabolic models describing the metabolism of breast, colon, and ovarian cancers. The *omics* data is used to define the topology and the physiology in the network. In particular, we used the metabolomics and the fluxomics to limit the corresponding metabolite concentrations and reaction rates in the model, and we used the transcriptomics data to define the set of transport reactions associated with each cancer type. Establishing the transport reactions in the model is of high importance, as these reactions will determine the exchanges with the medium setting at the same time the intracellular metabolism. We identified the transport reactions that are relevant for each cancer type based on the transcriptomics data and the network requirements. Furthermore, we used the cancer-specific models in combination with the transcriptomics data to infer the deregulation of the metabolic pathways by maximizing the consistency between the expression data and the allowable metabolic flux in the network. This method allows to identify the relation

between the deregulation of the genes and the deregulation of the metabolic pathways and to systematically compare the differences and similarities at the metabolic level across different types of cancers. The models have proven to be a valuable tool to assign deregulation to the reactions with a higher certitude than by just analyzing the expression data. Finally, we used the cancer-specific models to investigate the metabolic requirements for eleven well-known cancer phenotypes, including the Warburg effect, glutamine addiction, and cellular stress. For each phenotype, we assigned a set of metabolic tasks, and we identified all the alternative metabolic pathways that would need to be active in order to sustain each of the metabolic tasks. We performed enrichment analysis using the transcriptomics data, and we analyzed the pathways that were significantly deregulated in each phenotype of each cancer type. Our study highlights the differences in the metabolic deregulation for each cancer type and the models allow to give functionality to these deregulations based on the metabolic tasks that required the reactions that are catalyzed by the deregulated genes. The integration tools and the enrichment analyses we performed provide a deeper understanding of how cells adapt their metabolism and gene expression in the different cancer types. We thus provide one more degree of discrimination between cancers, which can ultimately lead to a better understanding of this disease and help in the design of personalized treatments.

In **Chapter 4**, we expand the metabolic models to integrate also the effects of the signaling processes controlling the expression of the metabolic genes. We extracted the human signal transduction pathways from Reactome, and we reconstructed the upstream signaling network for the transcription factor MYC. The signaling network contains the upstream pathways for MYC and the species that will determine the expression of the oncogene *MYC* [8]. We derived a novel method (CONSIGN) to contextualize the signaling network by maximizing the consistency of the states of its species with the observed data. We used the transcriptomics data form the breast cancer NCI60 cell lines to generate a signaling network for MYC in breast cancer cells. By analyzing the consistency of the states of the species with the data, we identified the proteins whose activity is required for the expression of MYC, and we analyzed the activity of the signaling pathways related to MYC in breast cancer, based on the consistent states of the species in the model. The models allow to infer the states of species that were not measured in the data based on consistency with the rest of the species in the network. Furthermore, we present a novel approach to integrate signaling

and metabolic networks by accounting for the regulation of the metabolic genes. We were able, for the first time, to simultaneously simulate the signaling and metabolic networks and the cross-talk between them. We integrated the breast-cancer specific MYC model and the metabolic breast cancer model derived in Chapter 2. The integrated model includes the regulatory effects of 30 transcription factors, that were part of the MYC signaling network, in the metabolic genes from the breast cancer model. We investigated the consistency of the integrated model with the data observing the effects that both biological networks have on each other. Now, that both metabolic and signaling processes have been intensively studied, the integrated models show promising applications to improve the predictions of the current models by including the information from both networks.

Throughout the work of this thesis, we enumerate the alternative routes that the metabolic and the signaling networks contain to represent the observed phenotype equally. These alternative pathways represent the flexibility of the cells to adapt to different environments by changing their expression profiles. Therefore, the enumeration and analysis of these alternatives are highly relevant for the study of drug targets as they can explain the resistance of cancer cells to therapies [9-12].

This thesis provides the models, methods, and approaches that will allow the scientific community to zoom-in in the biological processes of cells and analyze their alterations as cancer develops and progresses, as well as to characterize the metabolism and signaling pathways of the different species that populate the tumor microenvironment (Figure 5.1). The experimental data combined with the metabolic and signaling models and computational methods allow us to simulate metabolism and signaling pathways in a variety of cases, including the metabolism of tumor vs. normal cells, the metabolism of different cell types in the tumor microenvironment [13-15], such as normoxic cells vs. hypoxic cells [16] and the metabolism in immune cells vs. tumor cells to analyze how to engineer the immune cells to survive in the tumor microenvironment [17].

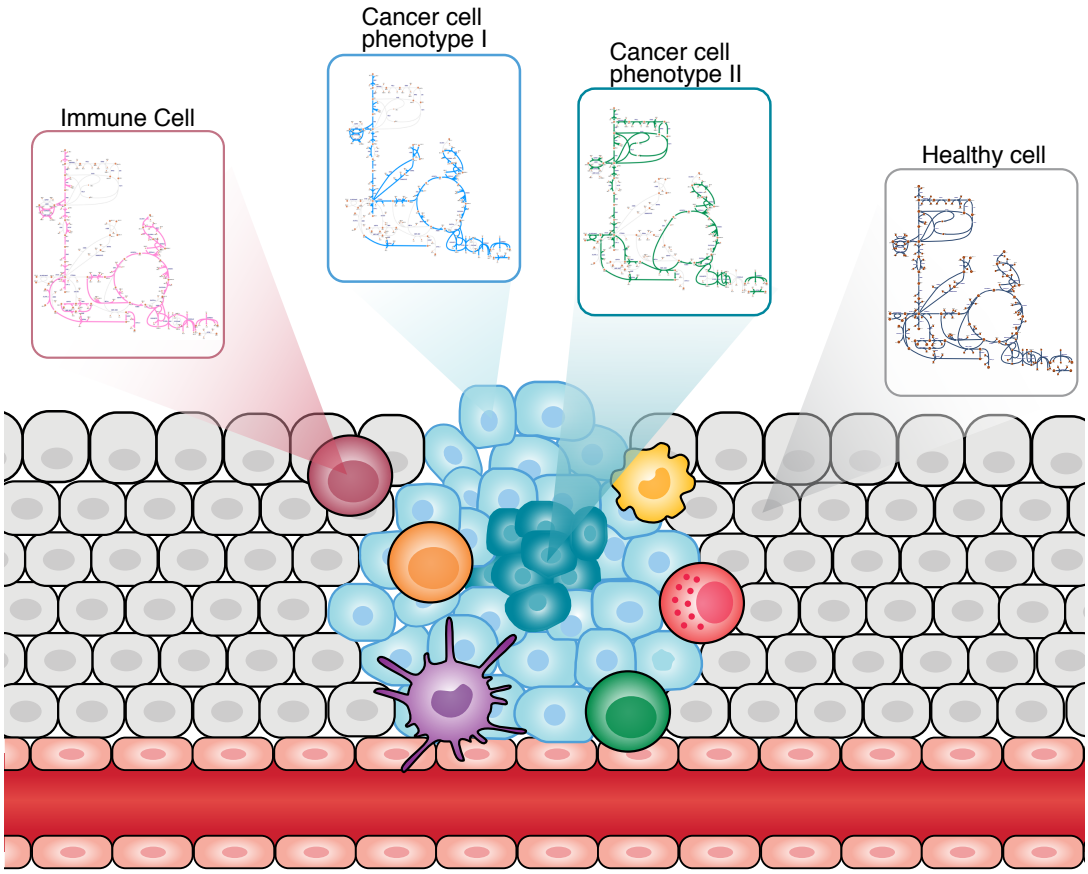


Figure 5.1. **Simulating the heterogeneity in the tumor microenvironment.** We have now the models and methods to simulate the behavior of the different type of cells that populate the tumor microenvironment allowing to investigate the similarities and differences among them. Figure adapted from [9].

5.2 Future perspectives

Systems Biology has evolved to simulate at the genome-scale the cellular processes occurring in the cells. However, the lack of data and complete knowledge of these processes required certain assumptions in order to build manageable models that could expand our understanding of the studied process. The models and methods derived allow us to interrogate the system and to generate a collection of hypothetical answers to biological questions.

Despite the assumptions made in this work, the models and methods generated can serve as a basis to build more detailed continuous models, as kinetic information and experimental data become available. The simpler models are used to explore the

structure of the system and enumerate a set of possible answers that can be later validated with improved versions of these models.

The models and approaches here developed can be immediately used to investigate the cellular phenotypes within the tumor. For this purpose, we would integrate *omics* data into the generic model, create metabolic and signaling models for the different types of cells, analyze their corresponding metabolic and signaling pathways within the tumor microenvironment and identify targets that strengthen the healthy cells and debilitate the tumor cells. Further validation of the generated models using experimental data would greatly improve their predictive capabilities, decreasing the uncertainty in the network and emphasizing the differences across cellular phenotypes. The current methods provide an upper bound with respect to the enzyme activity based on the expression profile. Further post-transcriptional and post-translational modifications such as phosphorylation, would further constrain the bounds of the activity of the enzymes, and therefore, future research directions should investigate how these post-transcriptional and post-translational modifications would change the bounds of these enzymes.

Genome-scale metabolic models and Boolean signaling models, in general, and those presented in this thesis, have shown remarkable applicability to infer the genotype-phenotype and to identify targets for drugs. These models enable us to simulate a specific state of the cell under particular conditions. However, they are very limited to simulate the evolution of the cell in time. The constraint-based models can serve as a scaffold to generate kinetic models [18-21] that are able to represent the kinetic mechanism of the system and simulate its dynamic behavior. We are currently developing a kinetic model for human metabolism using the cancer models derived in Chapter 3 and the framework ORACLE [21]. The Boolean models for signaling networks can be further improved by considering continuous constraints that allow to assign more states to the proteins than just active or inactive. We will integrate these constraints in the signaling model and in the integrated models derived in Chapter 3 to improve the consistency of the models with the integrated data and their predictive capabilities.

Furthermore, constraint-based models account only for the metabolic resources required to build the biomass building blocks that will allow the cell to grow and proliferate. However, cells need to use resources also to build the proteins and enzymes that will allow them to perform their functions. As an improvement of the thermodynamic flux

balance metabolic models, models that account also expression constraints, known as metabolic and expression models or ME-models [22-24], were created to simulate the allocation of resources for other cellular biological processes required for metabolism. Specifically, they integrate to the thermodynamically constraint-based model additional constraints to include enzyme and mRNA concentration levels. These models allow to improve the predictions of the metabolic model. We are currently adapting the ETFL workflow [22] to generate models for human metabolism that are able to simulate the transcription and translation of the metabolic genes in order to have enough enzymes to have activity in the metabolic pathways.

The integrated model is a powerful platform to investigate drug targets, drug effects and biomarkers. Simulating the interplay between signaling and metabolism can elucidate signaling and metabolic key components that could potentially decrease or stop the proliferative rate of cancer cells. The modularity of the model and the methods allows to integrate additional pathways that could represent the drug effect and the drug degradation pathways to study not only the impact of the drug but also the toxicity derived from its metabolism [25, 26]. In the future, we would like to test and validate the integrated breast cancer model generated in Chapter 3 (or expanded to contain other signaling pathways) against known drugs used for breast cancer [27, 28].

Finally, all the models and methods here generated have been used with cancer data from cultured cells. Nevertheless, other diseases showing altered phenotypes at the signaling and the metabolic levels could be studied following the same methodology. Moreover, using patient-specific data from the tumor and the healthy tissue around it could be used to create patient-specific models and help in the design of personalized medicine [29].

References

1. Thiele, I., et al., *A community-driven global reconstruction of human metabolism.* Nature Biotechnology, 2013. **31**(5): p. 419-+.
2. Brunk, E., et al., *Recon3D enables a three-dimensional view of gene variation in human metabolism.* Nature Biotechnology, 2018. **36**(3): p. 272-+.
3. Soh, K.C.a.H., V., *Constraining the flux space using thermodynamics and integration of metabolomics data.* Methods Mol Biol (Clifton, N.J.), 2014(1191): p. pp. 49-63.
4. Henry, C.S., L.J. Broadbelt, and V. Hatzimanikatis, *Thermodynamics-based metabolic flux analysis.* Biophysical Journal, 2007. **92**(5): p. 1792-1805.
5. Ataman, M., et al., *redGEM: Systematic reduction and analysis of genome-scale metabolic reconstructions for development of consistent core metabolic models.* Plos Computational Biology, 2017. **13**(7).
6. Ataman, M. and V. Hatzimanikatis, *lumpGEM: Systematic generation of subnetworks and elementally balanced lumped reactions for the biosynthesis of target metabolites.* Plos Computational Biology, 2017. **13**(7).
7. Jain, M., et al., *Metabolite Profiling Identifies a Key Role for Glycine in Rapid Cancer Cell Proliferation.* Science, 2012. **336**(6084): p. 1040-1044.
8. Dang, C.V., A. Le, and P. Gao, *MYC-induced cancer cell energy metabolism and therapeutic opportunities.* Clin Cancer Res, 2009. **15**(21): p. 6479-83.
9. Kreuzaler, P., et al., *Adapt and conquer: Metabolic flexibility in cancer growth, invasion and evasion.* Mol Metab, 2020. **33**: p. 83-101.
10. Vasan, N., J. Baselga, and D.M. Hyman, *A view on drug resistance in cancer.* Nature, 2019. **575**(7782): p. 299-309.
11. Gottesman, M.M., *Mechanisms of cancer drug resistance.* Annu Rev Med, 2002. **53**: p. 615-27.
12. Hawkes, N., *Drug resistance: the next target for cancer treatment.* BMJ, 2019. **365**: p. l2228.
13. Lyssiotis, C.A. and A.C. Kimmelman, *Metabolic Interactions in the Tumor Microenvironment.* Trends Cell Biol, 2017. **27**(11): p. 863-875.

14. Xiao, Z.T., Z.W. Dai, and J.W. Locasale, *Metabolic landscape of the tumor microenvironment at single cell resolution.* Nature Communications, 2019. **10**.

15. Hanahan, D. and L.M. Coussens, *Accessories to the Crime: Functions of Cells Recruited to the Tumor Microenvironment.* Cancer Cell, 2012. **21**(3): p. 309-322.

16. Eales, K.L., K.E. Hollinshead, and D.A. Tennant, *Hypoxia and metabolic adaptation of cancer cells.* Oncogenesis, 2016. **5**: p. e190.

17. O'Sullivan, D., et al., *Metabolic interventions in the immune response to cancer.* Nat Rev Immunol, 2019. **19**(5): p. 324-335.

18. Chakrabarti, A., et al., *Towards kinetic modeling of genome-scale metabolic networks without sacrificing stoichiometric, thermodynamic and physiological constraints.* Biotechnol J, 2013. **8**(9): p. 1043-57.

19. Wang, L. and V. Hatzimanikatis, *Metabolic engineering under uncertainty. I: framework development.* Metab Eng, 2006. **8**(2): p. 133-41.

20. Wang, L. and V. Hatzimanikatis, *Metabolic engineering under uncertainty--II: analysis of yeast metabolism.* Metab Eng, 2006. **8**(2): p. 142-59.

21. Miskovic, L. and V. Hatzimanikatis, *Production of biofuels and biochemicals: in need of an ORACLE.* Trends Biotechnol, 2010. **28**(8): p. 391-7.

22. Salvy, P. and V. Hatzimanikatis, *The ETFL formulation allows multi-omics integration in thermodynamics-compliant metabolism and expression models.* Nat Commun, 2020. **11**(1): p. 30.

23. O'Brien, E.J., et al., *Genome-scale models of metabolism and gene expression extend and refine growth phenotype prediction.* Molecular Systems Biology, 2013. **9**.

24. Lerman, J.A., et al., *In silico method for modelling metabolism and gene product expression at genome scale.* Nature Communications, 2012. **3**.

25. Bushweller, J.H., *Targeting transcription factors in cancer - from undruggable to reality.* Nat Rev Cancer, 2019. **19**(11): p. 611-624.

26. Felsher, D.W., *Cancer revoked: oncogenes as therapeutic targets.* Nat Rev Cancer, 2003. **3**(5): p. 375-80.

27. Niepel, M., et al., *Profiles of Basal and Stimulated Receptor Signaling Networks Predict Drug Response in Breast Cancer Lines.* Science Signaling, 2013. **6**(294).

28. Baselga, J., et al., *Pertuzumab plus Trastuzumab plus Docetaxel for Metastatic Breast Cancer.* New England Journal of Medicine, 2012. **366**(2): p. 109-119.

29. Chen, R. and M. Snyder, *Systems biology: personalized medicine for the future?* Curr Opin Pharmacol, 2012. **12**(5): p. 623-8.

MARIA MASID

Curriculum Vitae

Contact

Address:
Chemin de Chissiez 6
1006 Lausanne

Phone:
+34 646 854 156

Email:
maria.masid@gmail.com

Languages

Spanish – native
Galician – native
English – C1
French – B2

Skills

Microsoft Office, MATLAB, LaTeX, python (basic).

Team working, leadership, multidisciplinary work.

Summary

Mathematical background with experience in the computational study of human metabolism and its alterations in tumor cells. Strong background in multidisciplinary work environment, teaching and leadership.

Education

2016 – 2020 **PhD Student/Research Assistant**
École Polytechnique Fédérale de Lausanne, Switzerland

2012 – 2014 **Master in Mathematical Engineering**
University of Santiago de Compostela, Spain

2008 – 2012 **Bachelor of Science in Mathematics**
University of Santiago de Compostela, Spain

Research Experience

2016 – 2020 **École Polytechnique Fédérale de Lausanne, Switzerland**
PhD Studies. Supervisor: Prof. Vassily Hatzimanikatis
Topics:
- Mathematical modeling of human metabolism
- Cancer metabolic alterations
- Signaling and metabolic models

2014 – 2016 **University of Vigo, Spain**
Research assistant
Topic: Mathematical modeling and simulation of bone remodeling.

Teaching Experience

2019 – 2020 Supervising master projects
2016 – 2018 735h as teaching assistant at EPFL
2013 – 2014 Math tutoring high school and bachelor students

Publications

- **Masid, M.**, Pandey, V, et al. (2020). Contextualization of Signaling networks for the study of breast cancer (*in preparation)*
- **Masid, M.**, Hatzimanikatis, V. (2020). Model-based data integration and minimal network enrichment analysis identify metabolic differences across cancer types (*in preparation*).
- **Masid, M.**, Ataman, M., Hatzimanikatis, V. (2020). redHUMAN: analyzing human metabolism and growth media through systematic reductions of thermodynamically curated genome-scale models. Nat Commun 11, 2821 (2020).

- Fernández, J.R., Magaña, A., **Masid, M**., Quintanilla, R. (2019). Analysis for the strain gradient theory of porous thermoelasticity. Journal of Computational and Applied Mathematics, 345, 247-268.
- Fernández, J. R., **Masid, M**. (2018). A porous thermoviscoelastic mixture problem: numerical analysis and computational experiments. *Applicable Analysis*, 97:7, 1074-1093.
- Fernández, J.R., García-Aznar, J.M., **Masid, M**. (2017). Numerical analysis of an osteoconduction model arising in bone-implant integration. *ZAMM*. 97, 1050-1063.
- Fernández, J. R., **Masid, M**. (2017). Analysis of a model for the propagation of the ossification front. *Journal of Computational and Applied Mathematics*, 318, 624–633.
- Fernández, J. R., **Masid, M**. (2017) A porous thermoelastic problem with microtemperatures, *Journal of Thermal Stresses*, 40:2, 145-166.
- Fernández, J. R., **Masid, M**. (2017). A mixture of thermoelastic solids with two temperatures. *Computers and Mathematics with Applications*, 73(9), 1886-1899.
- Fernandez, J. R., **Masid, M**. (2017). Numerical analysis of a thermoelastic diffusion problem with voids. *International Journal of Numerical Analysis and Modeling*, 14(2), 153-174.
- Fernández, J. R., **Masid, M**. (2017). A porous thermoelastic problem: An a priori error analysis and computational experiments. *Applied Mathematics and Com putation*, 305, 117-135.
- Segade, A., Fernández, J. R., López-Campos, J. A., **Masid, M**., Vilán, J. A. (2017). A dynamic viscoelastic problem: Experimental and numerical results of a finite vibrating plate. *Cogent Mathematics*, 4(1), 1282691.
- Fernández, J. R., **Masid, M**. (2016) Analysis of a problem arising in porous thermoelasticity of type II, *Journal of Thermal Stresses*, 39:5, 513-531.
- *Copetti, M. I. M., Fernández, J. R., **Masid, M**. (2016). Numerical analysis of a viscoelastic mixture problem*. International Journal of Solids and Structures*, 80, 393-404.*

Conference Talks and Posters (last 3)

- Talk and Poster titled *Metabolic models enable the investigation of metabolic pathways deregulation in cancer cells, in 1st ECCSE Retreat in AGORA, Lausanne, Switzerland, 03 February 2020.*
- Talk titled *Metabolic models enable the investigation of metabolic pathways deregulation in cancer cells*, in Cancer SCCL Faculty and Staff retreat on 26-27 November 2019, at the Swiss Tech Convention Center at EPFL.
- Poster titled *Systematic reduction of genome-scale models for the study of metabolic phenotypes of human cells* presented at Constraint-Based Reconstruction and Analysis (COBRA) Conference Seattle, USA, 14-16 October 2018.

Grants

Early Stage Researcher Marie-Sklodoska Curie Fellowship in the ITN *SyMBioSys* during my PhD.