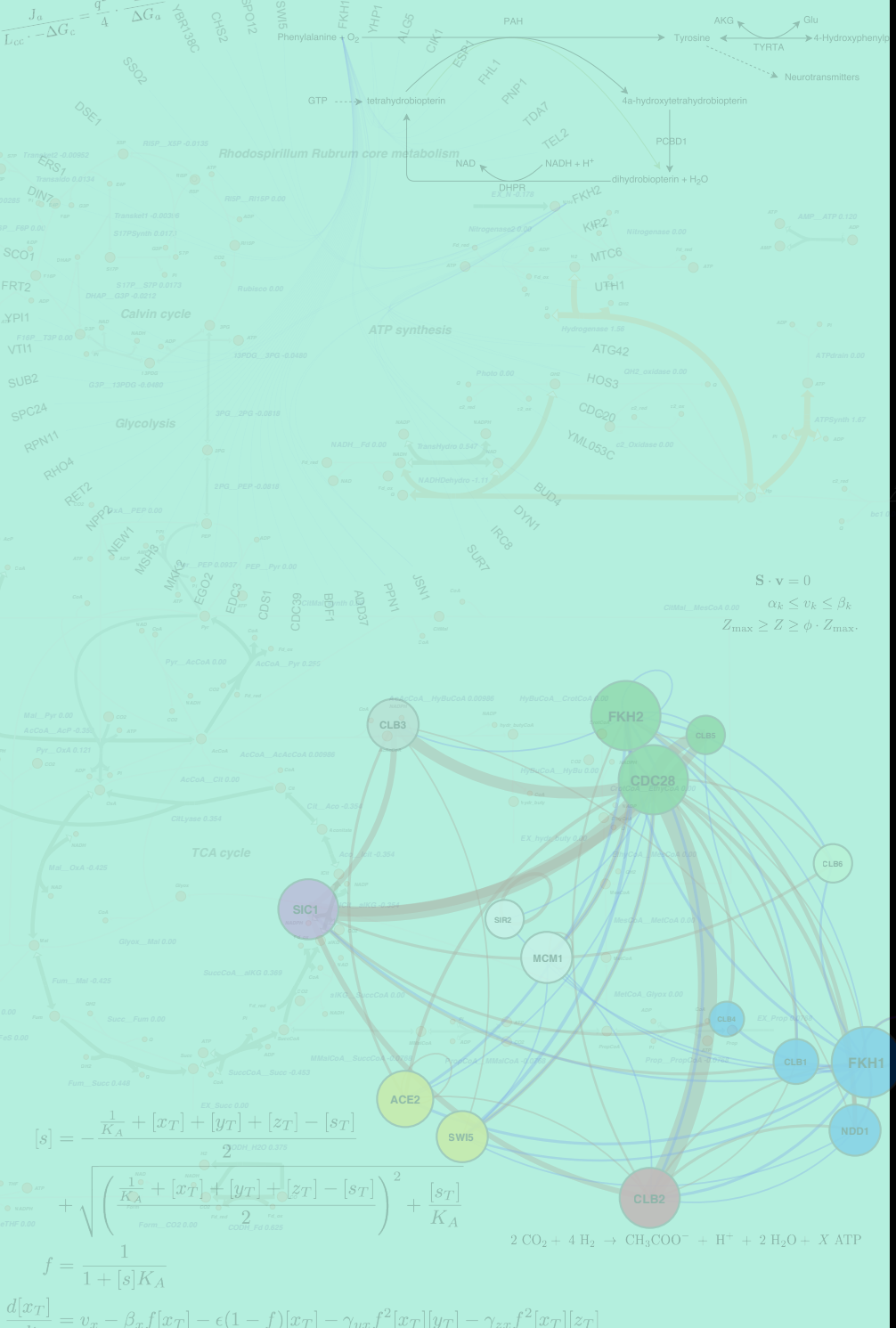
Waves, ChIPs, GEMMs, gears, markers and maps

Computational systems biology from cell cycle oscillations to metabolic fluxes

Mondeel, T.D.G.A.

Publication date
2022

Document Version
Final published version

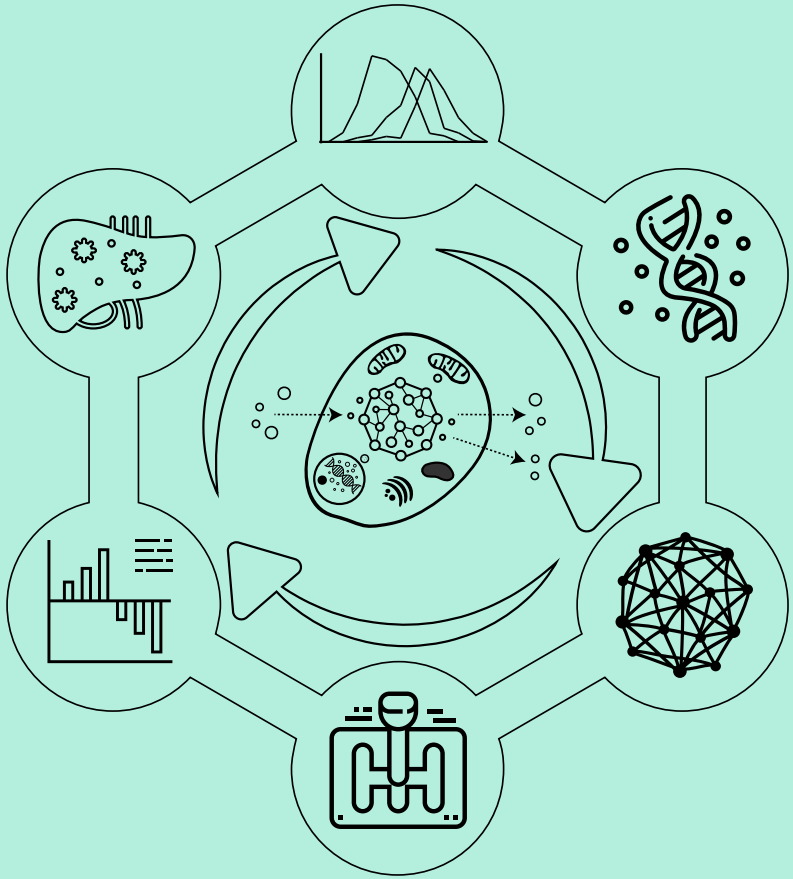


Waves, ChIPs, GEMMs, Gears, Markers and Maps

Thierry D.G.A. Mondeel

Waves, ChIPs, GEMMs, Gears, Markers and Maps

Computational Systems Biology from Cell Cycle Oscillations to Metabolic Fluxes



Thierry D.G.A. Mondeel

Waves, ChIPs, GEMMs, Gears, Markers and Maps
Computational Systems Biology
from Cell Cycle Oscillations
to Metabolic Fluxes

Thierry Dirk Gerrit Antoine Mondeel

The research described in this thesis was funded by the Research Priority Area Systems Biology of the University of Amsterdam.

The research was carried out within the group of Synthetic Systems Biology and Nuclear Organization, Swammerdam Institute for Life Sciences, University of Amsterdam, the Netherlands.

Waves, ChIPs, GEMMs, Gears, Markers and Maps
Computational Systems Biology from Cell Cycle Oscillations to Metabolic Fluxes

ACADEMISCH PROEFSCHRIFT

ter verkrijging van de graad van doctor
aan de Universiteit van Amsterdam
op gezag van de Rector Magnificus
prof. dr. ir. K.I.J. Maex
ten overstaan van een door het College voor Promoties ingestelde commissie,
in het openbaar te verdedigen in de Agnietenkapel
op woensdag 13 juli 2022, te 10.00 uur

door Dirk Gerrit Antoine Mondeel
geboren te Emmen

Promotiecommissie

Promotores:	prof. dr. H.V. Westerhoff	Universiteit van Amsterdam
	dr. M. Barberis	University of Surrey
Overige leden:	prof. dr. B.M. Bakker	Rijksuniversiteit Groningen
	dr. R.J. Tanaka	Imperial College London
	dr. ir. H.C.J. Hoefsloot	Universiteit van Amsterdam
	prof. dr. G. Muijzer	Universiteit van Amsterdam
	prof. dr. ing. A.H.C. van Kampen	AMC-UvA
	dr. R. Planqué	Vrije Universiteit Amsterdam

Faculteit der Natuurwetenschappen, Wiskunde en Informatica

To my parents and to Emma, for just about everything.

Contents

Contents **v**

1 General Introduction **1**
1.1 Beautiful, damn hard, and increasingly useful . . . 2
1.2 Networks, dynamics, systems, and objectives . . . 3
1.3 Putting the computation in systems biology . . . 4
1.4 From cell cycle oscillations to metabolic fluxes . . . 6

2 Clb3-centered regulations are recurrent across distinct parameter regions in minimal autonomous cell cycle oscillator designs **17**
2.1 Introduction . . . 18
2.2 Results . . . 22
Experimental rationale underlying the computational analyses . . . 22
The minimal cell cycle model and derivation of designs 1A–3 . . . 23
Conserved network motifs across oscillatory phenotypes . . . 25
Alternative network designs of the minimal cell cycle model . . . 27
The ability of extended designs to generate oscillations . . . 30
Limit cycles belong to distinct parameter space regions . . . 31
2.3 Discussion . . . 36
2.4 Methods . . . 41
Simulation of Ordinary Differential Equation (ODE) models . . . 42
Application of the SDS Toolbox . . . 42
Parameter space sampling to find oscillatory phenotypes . . . 42
Principal component analysis . . . 44
Supplementary Information . . . 46

3 ChIP-exo analysis highlights Fkh1 and Fkh2 transcription factors as hubs that integrate multi-scale networks in budding yeast **85**
3.1 Introduction . . . 86
3.2 Materials and Methods . . . 88
Yeast strains and growth conditions . . . 88
ChIP-exo . . . 88
Data processing . . . 89
Gene annotation and data analysis . . . 90
KEGG pathway map visualization . . . 91
3.3 Results . . . 91
Data analysis pipeline using the novel *maxPeak* method . . . 91
Consensus of verified and novel targets of Fkh1 and Fkh2 . . . 92
The correlation between Fkh and target expression levels . . . 96

Dynamics of cell cycle-regulated target genes . . . 97
Functional enrichment of identified Fkh target genes . . . 99
Fkh targets in their functional context . . . 101
3.4 Discussion . . . 103
Supplementary Materials and Methods . . . 111
Supplementary Text . . . 115

4 GEMMER: GEnome-wide tool for Multi-scale Modeling data Extraction and Representation for *Saccharomyces cerevisiae* 131
4.1 Introduction . . . 132
4.2 Features . . . 133
4.3 GEMMER's methodology . . . 133
4.4 Conclusions . . . 136

5 Gear-shifting across thermodynamic landscapes 139
5.1 Introduction . . . 141
The coupling of anabolism and catabolism . . . 142
Stochastic fluctuations, attractors and Onsager reciprocity . . . 143
The phenomenological stoichiometry . . . 152
5.2 Results . . . 156
NET works after all? Stability criteria . . . 156
The variomatic strategy . . . 157
Gear-shifting in acetogenic bacteria? . . . 158
The WL pathway and a hydrogenase are essential . . . 160
ATP coupled to the acetogenesis pathway . . . 161
BHB yield coupled to acetogenesis for plastic production . . . 165
Potential for gear shifting in *C. ljungdahlii* . . . 165
Gear-shifting in *S. sulfataricus* . . . 168
5.3 Discussion . . . 169
5.4 Methods . . . 170
Flux balance analysis . . . 170
Extending the *C. ljungdahlii* GeMM by Nagarajan et al. . . . 171
Model checking and visualization . . . 172
Reproducibility . . . 172

6 Flux balance analysis for biomarker prediction: A limited proof and in silico test for glutathione mediated drug-detoxification 175
6.1 Introduction . . . 176
6.2 Methods . . . 180
Biomarker prediction with flux variability analysis for IEMs . . . 180
Biomarker prediction for drug-metabolism with flux variability analysis . . . 182
Implementation and robustness of the FVA-based approach . . . 183
Kinetic biomarker prediction method and simulations . . . 183
Generating the FBA-capable Geenen et al. network . . . 184
Computational reproducibility . . . 184

6.3 Results . . . 184
Assessing the validity of the BPFVA method . . . 184
Kinetic biomarker predictions . . . 188
The oxoproline loop at the heart of the detoxification pathway . . . 190
Drug-induced metabolic changes . . . 192
Bypassing the oxoproline loop . . . 192
6.4 Discussion . . . 193
Supplementary Information . . . 197

7 Simultaneous integration of gene expression and nutrient availability for studying metabolism of hepatocellular carcinoma 219
7.1 Introduction . . . 220
7.2 Results . . . 222
GENSI methodology . . . 222
Metabolic genes: expression in two hepatoma cell lines . . . 223
Converting RNA-seq data to RAS . . . 227
Conversion of NA data into MUR . . . 227
An FBA-based scaling methodology . . . 228
Metabolic flux potential as predicted by flux variability analysis . . . 229
Experimental verification . . . 234
7.3 Discussion . . . 237
7.4 Materials and Methods . . . 239
Simulations . . . 243
In vitro experiments . . . 244
Supplementary Information . . . 246

8 General Discussion 257
8.1 Networks, coupling, dynamics and objectives . . . 257
8.2 Gear-shifting as a unifying lens . . . 262
8.3 The next wave: data, scalable tools, and integrated models of multi-scale biology . . . 263
8.4 The wisdom of the giants that came before . . . 266

Summary 269

Samenvatting 273

Acknowledgments 279

List of publications 285

CHAPTER 1

General Introduction

> "There is a theory which states that if ever anyone discovers exactly what the universe is for and why it is here, it will instantly disappear and be replaced by something even more bizarre and inexplicable. There is another which states that this has already happened."
>
> — Douglas Adams [1]

The phrase quoted above by English humorist and science fiction novelist Douglas Adams points to a universe that is bizarre and inexplicable. It also alludes to a purpose, a why, of the universe which may be discovered. Of course, the quote above is to be interpreted in its humorous context, but its subject matter extends into the realm of science. The questions: is there a why to the universe, what is that why, and does asking the 'why' question make sense when applied to the universe, are scientifically unanswered [2]. Nevertheless, these and other questions concerning complex components of this universe inspire scientists every day to keep looking for answers and discoveries across a wide variety of disciplines from cosmology to psychology to biology. Bypassing the elusive 'why' question, scientists typically try to gain more insight into the 'how' of the universe: How did the universe come about, how is it evolving, how does it maintain itself, and what is life? In doing so, even though the universe and our existence at times may appear quite bizarre and inexplicable, we assume a rational basis for the universe: we aim to discover principles, laws and theorems that provide the structure for us and the things around us.

In his 1944 book *What Is Life*, Nobel Prize-winning Austrian (quantum) physicist Erwin Schrödinger posed the question "How can the events *in space and time* which take place within the spatial boundary of a living organism be accounted for by physics and chemistry?" He concluded that "the obvious inability of present-day physics and chemistry to account for such events is no reason at all for doubting that they can be accounted for by those sciences" [3]. Indeed, the biological sciences have come a long way since 1944 and the ability of physics and chemistry, aided by mathematics and computer science, to account for biological events has grown tremendously.

The publication of Schrödinger's book inspired and was shortly followed by the discovery of the structure of DNA [4], and the field of molecular biology subsequently boomed. Simultaneously, a separate line of research with roots in non-equilibrium thermodynamics revealed that the coupling of multiple molecular processes is fundamental to life due to the second law of thermodynamics. This understanding gives rise to the need to study coupling between processes

and network properties [5]. The former line of research is more reductionist in nature, whereas the latter line of research is focused on whole systems and while merely phenomenological at its onset, soon incorporated mechanisms as well [6]. This thesis discusses several scientific works in the fields of systems biology and bioinformatics, which were formed from the convergence of the reductionist and systems ways of thinking on the occasion of the genomics revolution [7].

Systems biology continues in the spirit of Schrödinger's classic book but with a twist. Systems biology rephrases Schrödinger's question in terms of how biological function, i.e. life, emerges from the underlying molecular network which is constrained by physical and chemical laws. The twist unites the two roots of systems biology: we need to know both the components of the biological system and their properties, and how they interact and give rise to new functions and new top-down constraints. In the consideration of the whole network as a complex system, explicitly inclusive of molecular mechanisms, the fields of mathematics, systems theory and computer science have come to play prominent roles.

The works highlighted in this thesis illustrate (small) steps forward toward understanding the emergence of functions in biology from a systems perspective. The chapters that follow aim to contribute to answering Schrödinger's question, albeit without Schrödinger's sweeping scope. The scope of this thesis is narrower and emphasizes the use of mathematical modeling, computational approaches and bioinformatics to express and understand a small selection of the ideas and challenges found in modern biology. We will however return to Schrödinger's overarching question in the Discussion (Ch. 8).

1.1 Beautiful, damn hard, and increasingly useful

In a book of the same title and spirit as Schrödinger's, Margulis and Sagan illustratively suggested that life resembles a fractal, a pattern repeated when "zooming in" to smaller scales and when "zooming out" to larger scales [8]. In this sense, the fundamental unit or "fractal" of life is the cell, existing either as singular entities (e.g. bacteria and archaea), as communities of cells, or as the building blocks of multi-cellular organisms (e.g. humans). Viewed through the analogy of a fractal, the remark by mathematician Benoit Mandelbröt on fractals in mathematics also applies to the biology of life: "beautiful, damn hard, increasingly useful." The beauty and utility of life is evident around and through us. Biology being 'damn hard' is apparent from the mind-boggling complexity [9] and diversity [10] of single cells, the manner in which single cells unite to form communities [11] and multi-cellular organisms, and the fascinating way in which that complexity gives rise to the precise and coherent functioning evident throughout the natural world [12].

Underlying all the functionality and diversity of living organisms is the genome: an organism's entire set of long-term hereditary information. As proposed by Watson and Crick [4], molecularly, the genome consists of DNA, a double helix of two strands of nucleotide molecules held together by cross-strand hydrogen bonds. DNA in turn is compacted into individual packets called chro-

mosomes. In the same way that Morse code consists of dots interspersed with dashes, each chromosome is made up of functional, coding stretches of DNA, referred to as genes, interspersed between non-coding (but possibly functional) stretches of DNA [13]. Schrödinger's book inspired Watson and Crick toward their ultimate discovery of the structure of DNA, prompting Watson to write that Schrödinger "[. . .] very elegantly propounded the belief that genes were the key components of living cells and that, to understand what life is, we must know how genes act" [14].

Indeed, much of biological research since Watson and Crick has regarded the identification and functioning of genes. The complete human genome, which has since been sequenced, was initially estimated to house somewhere between 26,000-40,000 protein-coding genes [15, 16]. This number has been further downsized to below 25,000 genes [17] and recently to roughly 20,000 [18] as more complete drafts of the genome have been released. In contrast, the minimal number of genes needed to sustain (bacterial) life is currently thought to be ~ 400 based on gene knockout studies in the bacterium Mycoplasma genitalium [19]. These genes provide the blueprint for all cellular components: each gene encodes an RNA molecule which results from the transcription of a piece of DNA and is itself translated to produce proteins. Proteins in turn are three-dimensional molecules that perform a myriad of functions as catalysts, regulators of gene transcription and structural support units. The mind-boggling complexity of cells emerges from these thousands of functionally diverse, interacting components that exhibit distinct spatial and temporal properties. These interactions and the functions that emerge from them are what will concern us here.

1.2 Networks, dynamics, systems, and objectives

A reductionist approach may lead one to consider each organism as the sum of its cells and each cell as the sum of its parts. However, such a view is detrimental as it misses key aspects of the systems of life: emergence, complexity, responsiveness and robustness [20]. For complex systems, simply knowing the components of a system does not enable one to explain how functionality emerges, i.e. the whole is not just the sum of its parts: typically, the whole is a function of the parts. Therefore, to understand complex systems, it is crucial to understand (i) what the components can do and how they interact, (ii) the system's inputs and outputs and the relationships (coupling) between them, (iii) the dynamics across time of components and their interactions, and (iv) the system's goal or objective, if any, and how this objective is achieved through the interactions between the components. Following (i), the concept and science of "networks" is fundamental to modern biology [21]. Following (ii), we need to consider the network as a system that interacts with its environment. Following (iii), the mathematics of dynamical systems comes into play in biology. Finally, (iv) requires us to consider what cells are "wired" for, such as the generation of offspring. Is there an evolved objective and has there been a corresponding selection mechanism in the particular biological system under consideration?

Life, as such a complex system, is more like a verb: it repairs, maintains, reproduces and outdoes itself [8]. Furthermore, there is a fundamental need for systems biology to understand the cross-communication among biological pathways, due to the need to couple processes that require Gibbs energy input to processes that release Gibbs energy as a consequence of the second law of thermodynamics (see below and Ch. 5). The aim of systems biology is to discover general principles about how functional properties and behavior of living organisms arise from the interactions of their constituents, to understand these principles and systems and to predict their future states and behavior [22, 23]. These aims will underlie all chapters of this thesis. Specifically, all chapters (Ch. 2-7) deal with point (i) mentioned above, Ch. 2 and 6 deal with (iii), Ch. 3 deals with (iii) and (iv), and Ch. 5-7 deal with (ii) and (iv).

1.3 Putting the computation in systems biology

Schrödinger posited that the problem of life is a puzzle posed to no single discipline[1]. Given that cells consist of chemical compounds that are subject to the laws of physics, the relevance of chemistry and physics to cell biology is clear. Computer science, bioinformatics and mathematics play a role in modern biology for at least two reasons. First, the era of big data has reached biology [24, 25] as well as other parts of modern life [26]. Storage, retrieval and analysis of this data now frequently requires advanced computing architectures, bioinformatics pipelines and mathematical statistics. Ch. 3, 4 and 7 partially deal with large (although modest by today's new standards) datasets and networks. However, there is a second, more fundamental, reason for the importance of computer science, bioinformatics and mathematics in biology which arises directly from our mission to consider biological systems as complex systems as described above.

In order to understand the world through the scientific method, we must make observations and recognize patterns. This is what (big) data and their analysis can deliver. We also need something that (big) data itself cannot deliver: we need to construct hypotheses in the context of existing knowledge and proven principles. We can then predict and test the outcomes of our hypotheses and, if warranted, update the status of our hypothesis to that of a proven principle or fact. Unfortunately, for complex systems such as those of life, it is challenging to reason intuitively about thousands of components, each with a multitude of interactions with other components, which may change over time. If we want to transcend merely describing what we observe in the world and in the laboratory, exploratory and predictive modeling of some sort must be performed in order to support and go beyond our intuitive thinking.

The importance of predictive modeling for biology is amplified by our vast but limited ability to investigate biological systems experimentally. First of all, although technical capabilities do improve year by year, our ability to perform measurements in cells does not cover the entirety of the hypotheses we are capable of constructing. Even the recently introduced large scale -omics techniques,

[1] https://www.nature.com/articles/d41586-018-06166-x

such as transcriptomics and proteomics, (so far) fall short here because they are too variable, obtaining time courses is challenging and we are often unable to perform multiple experiments on a single biological system simultaneously (e.g. transcriptomics and proteomics measurements in the same cells at the same time). Secondly, there is a fundamental imprecision by which we can measure the real world, and there is fundamental stochasticity in molecular events and sample sizes. Both of these limitations can be remedied, at least to some extent, by computational analysis since it may nail down and suggest focused experimental tests. Furthermore, modeling is cost-effective, since experimental research is often more expensive. This is not to say that predictive modeling should or could replace experimental observation. Predictive modeling should instead serve as a compass, directing where to most effectively target experiments. It is for these reasons that mathematical analyses of experimental data and predictions, based on our (possibly flawed) understanding of biological systems, are highly valuable.

Pharmacologist James Black stated in his 1988 Nobel lecture: "Models in analytical pharmacology are not meant to be descriptions, pathetic descriptions, of nature; they are designed to be accurate descriptions of our pathetic thinking about nature. They are meant to expose assumptions, define expectations and help us to devise new tests." [27, 28]. Encapsulating our current understanding of a system (or parts thereof) in a model and precisely formulating how we think the components interact with each other and the environment enables us to make precise predictions. These predictions are falsifiable when they lie within the sphere of experimentally obtainable results. Not only can such theoretical predictions lay bare naive conceptions in our understanding of the system, they can highlight exactly how specific design principles emerge and give rise to their ultimate effects, and may suggest new, focused experiments to perform. Iteratively improving our understanding of biological systems in this way can be seen as an upward (in the sense of understanding) spiral of systems biology [23, 29]. Sir Arthur Conan Doyle wrote in the Adventure of Sherlock Holmes, "Data! Data! I can't make bricks without clay!" [30]. A systems biologist might rephrase this statement and say: "Data! Model! Data! Model! I can't understand life without data and a model!"

The explosion of modeling in biology has led to the development and application of many modeling formalisms and tools [31]. Two techniques will be applied extensively in this thesis. Dynamic models using ordinary differential equations [32] will be featured in Ch. 2 and 6. Such models describe the temporal trajectories of concentrations and fluxes of components of a biological system in terms of kinetic and thermodynamic properties of catalytic or information components such as enzymes, transporters and transcription factors. The differential equations may be equated to zero and solved to find and analyze the properties of their steady state(s) such as stability (in terms of eigenvalues) and metabolite concentrations. Steady states are points in the phase space of the system (the space given by the dependent variables) where the dynamics of the system stay put, i.e. the rates of change of all variables in the system are zero. As discussed in Ch. 2 and 6, steady states may reflect biological states that are functional (healthy)

or dysfunctional (diseased), and many interesting properties may be calculated from them [32, 33]. The second technique, flux balance analysis [34], a form of linear programming, will be featured in Ch. 5-7. In contrast to ordinary differential equation models, flux balance analysis models do not deal with time and concentrations as such but consider the system at steady state and solely in terms of a flux pattern. Instead of temporal evolution of the system they describe the flow (or flux) through a biological network at steady state.

Ch. 2 and 5-7 use mathematical modeling to interpret observed biological phenomena, and where possible, suggest new experiments to test model predictions. Furthermore, Ch. 3 and 7 present new experimental datasets, their computational and bioinformatics analysis, ramifications and subsequent testable predictions. Ch. 4 discusses a new web-based tool for making sense of existing genome-wide data in the literature. Finally, Ch. 7 applies new methods for incorporating transcriptome data and medium-metabolite concentrations into flux balance analysis.

1.4 Computational systems biology from cell cycle oscillations to metabolic fluxes

Chapters 2-7 present mathematical modeling and data analyses on a wide variety of topics: cyclin/Cdk1 oscillations during the yeast cell cycle modeled using ordinary differential equations (Ch. 2); data analysis, integration and visualization (Ch. 3-4); steady state metabolic flux predictions in acetogenic bacteria (Ch. 5); human liver in the context of drug detoxification (Ch. 6) and human hepatocellular carcinoma cell lines (Ch. 7). However, as divergent as the chapters are in terms of their biological focal points and in terms of the computational techniques applied, at the heart of each is the primacy of the network, its stationary states, and the integrated behavior of its components as a system engaging in various modes of achieving its function. Ch. 2-7 can be grouped into 3 parts based on the biological topic under consideration: (i) the process during which life creates a replica of itself: *the cell cycle* (Ch. 2-3), (ii) visualizing and presenting the vast amount of information available on the protein-protein interaction network in budding yeast (Ch. 4), and (iii) the processes through which cells build more of themselves by building new components using nutrients from their environments: *metabolism* (Ch. 5-7). Ch. 4 in a way lives in between Ch. 2-3 and Ch. 5-7 in that it considers interaction networks in a genome-wide sense, not just those nodes relevant to one particular biological process. Ch. 5-7 will often assume that the system aims to optimize biomass production which, similar to Ch. 2 and 3, also involves life generating more of itself. Fig. 1.1 illustrates the grouping and content of the chapters. Ch. 3 and 4 are focused on data analysis and data visualization whereas Ch. 2 and Ch. 5-7 use different modeling techniques but all are focused on steady state related behavior.

To lighten the scientific jargon and make the content of this thesis more approachable, the topics of the following six chapters were summarized in the title of this thesis through the terms: *waves*, *ChIPs*, *GEMMs*, *gears*, *markers* and *maps*.

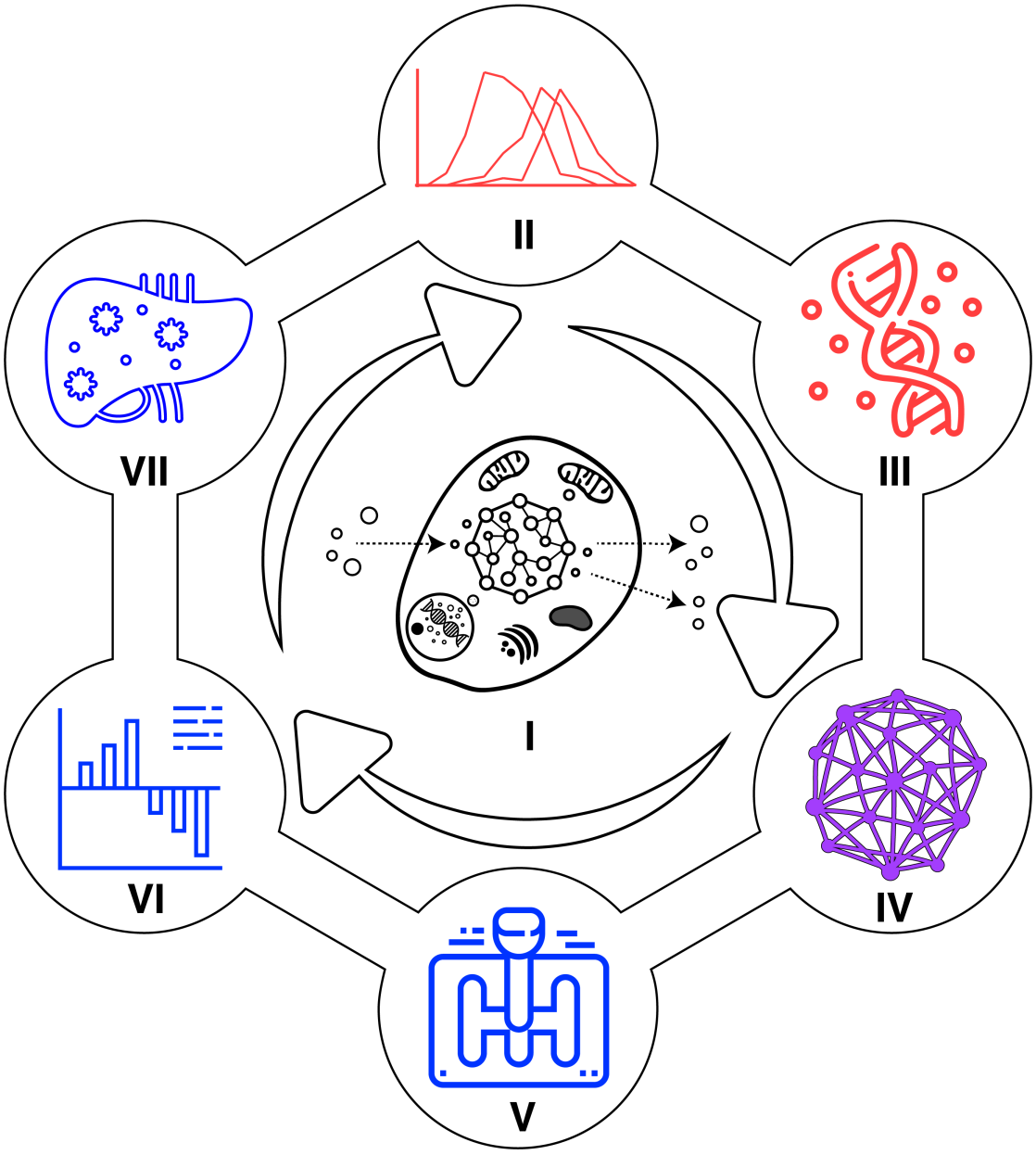


Figure 1.1: Overview of the six interrelated scientific topics dealt with in this thesis. Ch. 2 and 3 are generally concerned with the budding yeast cell cycle (red), where Ch. 2 is concerned with oscillatory phenomena in the cell cycle regulatory network and Ch. 3 with DNA binding of two pivotal transcriptional regulators in that network. Ch. 5-7 relate to metabolism (blue), with Ch. 5 placing an emphasis on thermodynamics and energy harvesting in prokaryotes, and Ch. 6 and 7 focus on biomarker prediction and metabolic shifts in response to nutrition changes respectively. Ch. 4 (purple) bridges the gap from Ch. 2-3 to Ch. 5-7, providing a logical connection between them, and is focused on the organization, representation and visualization of the knowledge of gene and protein interactions across the whole genome for budding yeast.

These respectively refer to: (i) *waves*, or oscillations, of cyclin/Cdk1 complexes throughout the budding yeast cell cycle (Ch. 2), (ii) chromatin immunoprecipitation with exonuclease treatment experiments (*ChIP*-exo) to assess the genomic binding locations of the forkhead transcription factors in budding yeast (Ch. 3), (iii) the web-based network visualization tool *GEMMER* that we developed for genetic and protein-protein interaction networks in budding yeast (Ch. 4), (iv) the concept of *gear-shifting* and metabolic optimality discussed in (Ch. 5), (v) the prediction of *(bio)markers* of glutathione conjugation in the liver from genome-wide metabolic maps and from a kinetic model (Ch. 6), and (vi) GEnome-scale Metabolic (*GEM*) *maps* which are utilized in Ch. 5-7.

Each chapter includes a stand-alone introduction and discussion. Below, the content of each chapter is briefly introduced from a more general perspective and in the context of the topics (networks, dynamics, systems, computation and thermodynamics) and terms (waves, ChIPs, GEMMs, gears, markers and maps) discussed above.

Waves: first autonomous oscillator model for cyclin/Cdk1 oscillations in the budding yeast cell cycle

Life, from bacteria to humans, grows and renews through cell division, the process in which a mother cell divides into two daughter cells. In eukaryotes, progression through the cell cycle is associated with oscillations of complexes between cyclins and cyclin-dependent kinase molecules. The network giving rise to these oscillations has a fundamental design that is common to different organisms [35]. Budding yeast has been used as a model organism to study cell cycle regulation [36]. In Ch. 2 we use dynamical systems theory [37] to better understand the network properties that enable the budding yeast cyclin/Cdk1 network to exhibit steady oscillations, without being driven or reset by an external force.

Functional and dysfunctional cellular states may be viewed as (different) stable attractors, i.e. some alteration has been introduced in the functional biochemical network or environment that gives rise to a different attractor that is dysfunctional [38]. These attractors are typically stationary states in a high-dimensional state space (also called the phase space): they tend to attract systems that are in nearby states and are fixed once entered. The state dimensionality here is equal to the number of independent time-varying quantities such as molecular concentrations within the cell and in the environment. The point in state space that a system assumes at steady state is determined by a great many parameters that together span a parameter space. The dimensionality of and the location in this parameter space is mostly determined by the genome, by physical and chemical constraints (e.g. thermodynamics and diffusion constraints) and by the environment. Often stationary states are steady, i.e the point in phase space does not move with time. For oscillating systems such as circadian rhythms and the cell cycle, the attractor may be a stationary state that changes position in state space with time, but returns to an earlier position at some point in time and then repeats its behavior. This is the case of stable oscillations (also called 'limit cycle oscillations') where the system moves through a repeating continuum of states

that is fixed. A third type of system - i.e. that of deterministic chaos - continues to move through the state space, never turns back upon itself, but remains within a confined area of state space. A fourth type of system does not remain within any bound; these are globally unstable systems, impossible in biology, because biology is always constrained by resources.

In this thesis we shall only consider the first two types of stationary systems and in particular those that are stable in the sense of Lyapunov, i.e. that, after a small perturbation in variable values, return asymptotically to the limit of the same stationary state. A stable oscillation will thereby persist even in the face of natural perturbations, and it is then called a stable limit cycle.

The state space may have multiple attractors for the same set of parameter values, and the functional and dysfunctional states may correspond to different attractors for the same parameter values. In this case, the disease may be caused by a transient perturbation of variable values followed by a relaxation to a different attractor. The state space might also have a single attractor and the disease is then caused by a steady change in parameter value, which shifts the attractor in state space. The latter case may correspond to a genetic disease.

In Ch. 2 we focus on the budding yeast cell cycle in terms of a stationary state of stable oscillations in (complexes of) four proteins that play a role in regulating cell cycle progression, starting from a minimal network model generated in our laboratory [39, 40]. Ch. 2 mainly addresses the questions: (i) Can we construct a mathematical model, true to biological knowledge about the system, that oscillates in the sense of being a stable limit cycle? (ii) Are there substantial differences among different hypothetical and known network designs in budding yeast that are particularly able to yield oscillating dynamics corresponding to the cell cycle? (iii) Are there multiple distinct areas in the parameter space where these oscillations may occur, or is there just a single one? The goal here is to gain insight into design principles of the cell cycle regulatory network that allow it to produce oscillations, and to ascertain whether there are multiple ways (different parameter settings) of doing so. This is a task ideally suited to modeling-based analysis. In Ch. 2, instead of taking a classical continuation approach to bifurcation analysis, we use a recently proposed modeling methodology [41] which breaks any kinetic model up into a distinct set of *phenotypes* with particular properties. We utilize this approach by sampling the set of phenotypes for our models to identify and characterize points in the parameter and state space where the system oscillates. This results in a collection of parameter sets that may lie in vastly different areas of the parameter space and produce stationary cell cycle oscillations with different properties. This collection of parameter sets is then analyzed for conserved features and patterns.

ChIPs: ChIP-exonuclease analysis of Forkhead transcription factors in budding yeast

Continuing from Ch. 2, Ch. 3 zooms in on the Forkhead transcription factors, Fkh1 and Fkh2, which are members of a family of transcription factors that is conserved among eukaryotes [42]. Fkh1 and Fkh2 are responsible for several of

the interactions in the networks we studied in Ch. 2, in terms of their involvement in the generation of cell cycle oscillations.

In order to respond accurately to the cell's internal and environmental cues, the cell cycle must in some way interact with other cellular processes such as signal transduction and metabolism. Both Fkh1 and Fkh2 are intriguing candidates for the connections between the cell cycle, signal transduction and metabolism because they: (i) have relatively lengthy expression windows, (ii) are known to play crucial roles in the cell cycle regulatory network [40] and (iii) may target [43–45] more than just cell cycle genes. Therefore, in Ch. 3 we investigate the question: Can the target genes of Forkhead transcription factors Fkh1 and Fkh2 play roles in processes other than the cell cycle? If so, they could function as hubs connecting multiple cellular processes.

We approach this question by performing the so-called ChIP-exo [46], a relatively recent experimental approach that can detect protein-DNA interactions at near single-nucleotide resolution, for Fkh1 and Fkh2. We shall do this across two cellular conditions, logarithmic phase (i.e. exponential growth) and stationary phase caused by nutrient depletion. We will match the genomic binding locations of Fkh1 and Fkh2 to the promoters of genes and, consequently, gain insight into the genes they may regulate within and outside the cell cycle. In the course of our analysis we make use of two existing peak detection tools and develop a new, perhaps more appropriate method: *maxPeak*. In addition to analyzing the resulting data, we extensively catalogue how the new data matches and updates previously available data for Fkh1 and Fkh2.

GEMMs: Visualizing interaction networks in budding yeast

Continuing on the theme of integrating cellular processes, Ch. 4 turns to visualizing the available data on gene-gene, protein-gene and protein-protein interactions for budding yeast, across space, time and functional scales. We perceived that there was a need to integrate various separate databases containing information on interactions alone, spatial characteristics alone and timing alone while projecting this information onto interaction networks. This should then allow interaction networks to be viewed in their spatial, temporal and functional contexts. To this purpose, we built 'GEMMER', a web-based tool that allows for such visualizations to be produced in a user-friendly manner. GEMMER should aid researchers to "rediscover" well-known interactions in their full genome-scale context, and perhaps to spot lesser-known interactions that may be of relevance to certain scientific questions.

Gears: Gear-shifting in early micro-organisms

After discussing the budding yeast cell cycle (Ch. 2 and Ch. 3) and integrating various alternative cellular processes (Ch. 3 and Ch. 4), Ch. 5 turns to another fundamental cellular process: metabolism. Metabolism refers to the processes through which cells build their own components (anabolism) and break down nutrients from their environments (catabolism).

As discussed in Ch. 2 and 3, the cyclin-dependent kinase Cdk1 progressively binds to different cyclins throughout the phases of the cell cycle. By doing this, cyclin subunits shift gears in the sense of redirecting Cdk1 between proteins that thereby become subject to phosphorylation. In microorganisms, certain proteins which alternate partner proteins also cause changes between similar functionalities. This is what we identify as 'shifting gears'. The organism may also shift between different flux patterns through its metabolic network, a third form of 'gear shifting'.

Consideration of life as a verb, i.e. constituted by cells that actively change, divide and respond to stimuli, directly ties into their need to transform forms of energy. It thereby connects with one of the foundational fields of physics: thermodynamics. Thermodynamics proposes that the course of each chemical event in the universe is dictated by the energy content of the system under consideration and the energy exchange between the system and its environment [47], where this energy can have various forms.

Organisms are like islands of order in an ocean of chaos [8]. In order to meet the second law of thermodynamics, as the order inside an organism increases, the disorder in the universe as a whole must increase; for processes and life to happen chaos must be produced in the universe as a whole, or, more stringently, Gibbs or metabolic energy must be dissipated. The maintenance of the order inside an organism, and the continual need for synthesis and replacement of the chemical components making up the cell require continuous Gibbs (or metabolic) energy input [6], and a coupling of processes that lower Gibbs energy to the processes that require Gibbs energy input. Thermodynamics furthermore dictates the direction in which chemical reactions occur and as such poses fundamental restrictions on network performance. In biology, evolution has led to energy coupling, whereby the direction of chemical reactions can be inverted through coupling to other reactions that are downhill in terms of Gibbs energy.

In Ch. 5 we deal with both equilibrium thermodynamics (ET) and non-equilibrium thermodynamics (NET) in the context of metabolism of prokaryotes that may have populated earth in its early days. ET, dealing with the initial and final states of the system, is not concerned with time, rates or the pathway by which a chemical change occurs. It merely reasons based on the change in Gibbs or metabolic energy between the initial and final states of the system. In contrast, NET does deal with time, rates, coupling, yields and actual thermodynamic efficiency [6].

At equal catalytic efficacy, flux (rate) increases with a process's Gibbs energy dissipation: i.e. the more thermodynamically favorable a process is, the higher the flux rate tends to be [48]. In Ch. 5, we discuss a variomatic gear-shifting principle, by way of coupling anabolism and catabolism, illustrating how cells may increase anabolic flux rates (e.g. growth rate) by changing the degree of coupling between anabolism and catabolism, not by traditionally decreasing the 'leakage' or 'slippage' but by increasing the stoichiometry. We further consider whether micro-organisms are capable of achieving gear-shifting through the proteins (e.g. enzymes) they express such that they optimize flux yield. Gear-shifting may be a strategy for organisms to regulate the trade-off between rate and yield. By se-

lectively expressing only certain enzymes, the amount of ATP synthesis coupled to a process may increase, thereby reducing the thermodynamic favorability of the overall process and decreasing the rate. Schüchmann and Müller posited that acetogenic bacteria, and *Cl. ljundahlii* in particular, may be able to couple different amounts of ATP synthesis to the process of acetogenesis depending on the electron acceptors/donors of the hydrogenase and the enzymes of the Wood-Ljungdahl Pathway (WLP) [49]: i.e. for the same input and output, a different ATP yield may be achieved by shifting between enzymes in the pathways involved. Using flux balance analysis simulations, we highlight that indeed *Cl. ljungdahlii* is at least in theory able to gear-shift in the acetogenesis process.

Markers: illustrating biomarker prediction for glutathione conjugation: steady state metabolic maps versus dynamic models

Sticking with the metabolic theme, in Ch. 6 we move from prokaryotic to human metabolism and continue with the application of flux balance analysis to metabolic maps. A key promise of the advent of the human metabolic map [50, 51] was the ability to more systematically investigate the causes, consequences and monitoring of diseases. Flux balance analysis only addresses fluxes and will not help predict changes in concentrations indicative of ('biomarking') disease. Following the first release of the human metabolic map, a computational approach using flux variability analysis (FVA) was proposed to predict concentration changes that could serve as biomarkers of human inborn errors of metabolism. This approach appeared to produce fair agreement with experimentally validated sets of biomarkers [51, 52]. In Ch. 6 we focus on this method of biomarker prediction. We first develop a hitherto lacking rationale for the approach, then re-implement it, double-check its reported performance, and then apply it to a search for biomarkers of glutathione conjugation capacity in the human liver.

Before supplying patients with xenobiotics that deplete glutathione, it would be advisable to infer the patient's glutathione status in the liver by way of measuring blood biomarkers. In Ch. 6, we investigate whether the FVA-based biomarker prediction methodology recovers the suitability of serum ophthalmic acid (OPA) and 5-oxoproline (OXO) as robust biomarkers of glutathione levels and utilization, as predicted by a kinetic model of the same system that is used as the gold standard.

Maps: nutritional shifts and the Warburg effect in two liver cancer cell lines

Metabolism is directly involved in many human diseases including cancer, and indirectly in virtually all, because disease causes metabolic changes that can accompany or even affect etiologies and be read as biomarkers. The best-known metabolic abnormality in cancer cells is an increased glycolysis followed by lactic acid production even in the presence of oxygen and fully functional mitochondria, a process known as the Warburg Effect [53]. It is as if cancer cells perma-

nently shift to a lower gear in terms of diminished ATP production from glucose. Ch. 7 further expands on the theme of metabolism in human disease states and gear-shifting by turning to the application of flux balance analysis to the human metabolic reconstruction [54] in the context of hepatocellular carcinoma. We combine this genome-scale metabolic modeling approach with in vitro experiments to investigate whether the behavior of cancer cells is determined by their nutrition or/and the expression of their genes. To that end, we developed a new method that integrates the genome (nature) and the environment (nurture) and identifies the influence of cell-nutrition changes on the Warburg effect in two hepatocellular carcinoma cell lines.

In Ch. 7 we use flux balance analysis to investigate whether two hepatocellular carcinoma cell lines show metabolic differences, particularly in terms of their Warburg effect, when cultured in media with various carbon and Gibbs energy sources. Experimentally, we characterize the cells using exometabolomics and growth rate assays. Computationally, we propose a novel extension of flux balance analysis by which we predict metabolic flux patterns consistent with both the measured transcriptome profiles of the cell lines, through transcriptome-limited flux bounds, and specific medium conditions in the flux balance analysis models. The medium conditions are represented by their experimentally measured metabolite concentrations which will be translated into flux bounds in the flux balance simulations. The integration of these two types of constraints requires a choice of a scaling constant weighing the two sets of constraints. By choosing this scaling constant carefully, a model can be generated that closely matches observed experimental behavior.

Further data, technique and tool development, and the future of big data and big models in biology

This thesis concludes in Ch. 8 with a general discussion on the relevance of the themes developed in chapters 2-7 and their connections for future biomedical research. Particularly, we discuss the importance of the biological principles discovered and further development of the novel data (Ch. 3 and 7), techniques (Ch. 2, 5, 6 and 7) and tools (Ch. 4) discussed in this thesis. Finally, we also view the work in chapters 2-7 in terms of the wider aspects of the big data and big model explosion in biology which will continue to have strong impact in biomedical research.

References

[1] D. Adams. *The Restaurant at the End of the Universe: Hitchhiker's Guide 2*. Vol. 2. Tor UK, 1989.

[2] S. M. Carroll. "Why Is There Something, Rather Than Nothing?" (2018). 10.48550/arXiv.1802.02231.

[3] E. Schrödinger. *What is life?* University Press: Cambridge, 1944.

[4] J. D. Watson and F. H. C. Crick. "Molecular Structure of Nucleic Acids: A Structure for Deoxyribose Nucleic Acid". *Nature* 171 (1953), pp. 737–738. 10.1038/171737a0.

[5] O. Kedem and S. R. Caplan. "Degree of coupling and its relation to efficiency of energy conversion". *Transactions of the Faraday Society* 61 (1965), p. 1897. 10.1039/tf9656101897.

[6] H. V. Westerhoff and K. van Dam. *Thermodynamics and control of biological free-energy transduction*. Amsterdam: Elsevier, 1987. 10.15490/fairdomhub.1.datafile.4954.1.

[7] H. V. Westerhoff and B. O. Palsson. "The evolution of molecular biology into systems biology". *Nature Biotechnology* 22 (2004), pp. 1249–1252. 10.1038/nbt1020.

[8] L. Margulis and D. Sagan. *What is life?* Univ of California Press, 2000.

[9] J. R. Karr *et al.* "A Whole-Cell Computational Model Predicts Phenotype from Genotype". *Cell* 150 (2012), pp. 389–401. 10.1016/j.cell.2012.05.044.

[10] S. Patange, M. Girvan, and D. R. Larson. "Single-cell systems biology: Probing the basic unit of information flow". *Current Opinion in Systems Biology* 8 (2018), pp. 7–15. 10.1016/j.coisb.2017.11.011.

[11] R. M. Stubbendieck, C. Vargas-Bautista, and P. D. Straight. "Bacterial Communities: Interactions to Scale". *Frontiers in Microbiology* 7 (2016), pp. 1–19. 10.3389/fmicb.2016.01234.

[12] V. V. Isaeva. "Self-organization in biological systems". *Biology Bulletin* 39 (2012), pp. 110–118. 10.1134/S1062359012020069.

[13] E. Pennisi. "ENCODE Project Writes Eulogy for Junk DNA". *Science* 337 (2012), pp. 1159–1161. 10.1126/science.337.6099.1159.

[14] J. D. Watson. *The Double Helix: A Personal Account of the Discovery of the Structure of DNA*. Atheneum, 1968.

[15] E. S. L. Lander *et al.* "Initial sequencing and analysis of the human genome". *Nature* 409 (2001), pp. 860–921. 10.1038/35057062.

[16] J. C. Venter *et al.* "The Sequence of the Human Genome". *Science* 291 (2001), pp. 1304–1351. 10.1126/science.1058040.

[17] International Human Genome Sequencing Consortium. "Finishing the euchromatic sequence of the human genome." *Nature* 431 (2004), pp. 931–45. 10.1038/nature03001.

[18] S. L. Salzberg. "Open questions: How many genes do we have?" *BMC Biology* 16 (2018), p. 94. 10.1186/s12915-018-0564-x.

[19] J. I. Glass *et al.* "Essential genes of a minimal bacterium". *Proceedings of the National Academy of Sciences* 103 (2006), pp. 425–430. 10.1073/pnas.0510013103.

[20] M. H. V. V. Regenmortel. "Reductionism and complexity in molecular biology". *EMBO Reports* 5 (2004), pp. 6–10.

[21] A.-L. Barabási and Z. N. Oltvai. "Network biology: understanding the cell's functional organization". *Nature Reviews Genetics* 5 (2004), pp. 101–113. 10.1038/nrg1272.

[22] L. Alberghina and H. V. Westerhoff. *Systems biology: definitions and perspectives*. Vol. 13. Springer Science & Business Media, 2007.

[23] H. V. Westerhoff and D. B. Kell. "The methodologies of systems biology". *Systems Biology*. Elsevier, 2007, pp. 23–70. 10.1016/B978-044452085-2/50004-8.

[24] F. S. Collins. "The Human Genome Project: Lessons from Large-Scale Biology". *Science* 300 (2003), pp. 286–290. 10.1126/science.1084564.

[25] V. Marx. "The big challenges of big data". *Nature* 498 (2013), pp. 255–260. 10.1038/498255a.

[26] D. Boyd and K. Crawford. "Critical questions for big data: Provocations for a cultural, technological, and scholarly phenomenon". *Information Communication and Society* 15 (2012), pp. 662–679. 10.1080/1369118X.2012.678878.

[27] J. Black. "Drugs from emasculated hormones: the principle of syntopic antagonism". *Science* 245 (1989), pp. 486–493. 10.1126/science.2569237.

[28] J. Gunawardena. "Models in biology: 'accurate descriptions of our pathetic thinking'". *BMC Biology* 12 (2014), p. 29. 10.1186/1741-7007-12-29.

[29] H. Kitano. "Computational systems biology". *Nature* 420 (2002), pp. 206–210. 10.1038/nature01254.

[30] A. Conan Doyle. *The Adventures of Sherlock Holmes*. London, United Kingdom: George Newnes, 1892, p. 307.

[31] D. Machado *et al.* "Modeling formalisms in Systems Biology". *AMB Express* 1 (2011), p. 45. 10.1186/2191-0855-1-45.

[32] J. J. Tyson, K. C. Chen, and B. Novak. "Sniffers, buzzers, toggles and blinkers: dynamics of regulatory and signaling pathways in the cell". *Current Opinion in Cell Biology* 15 (2003), pp. 221–231. 10.1016/S0955-0674(03)00017-6.

[33] H. Kacser, J. A. Burns, and D. A. Fell. "The Control of Flux: 21 Years On The control of flux". *Pharmaceuticals* (1995), pp. 18–1995.

[34] A. Varma and B. O. Palsson. "Metabolic Flux Balancing: Basic Concepts, Scientific and Practical Use". *Nature Biotechnology* 12 (1994), pp. 994–998. 10.1038/nbt1094-994.

[35] A. Csikász-Nagy *et al.* "Analysis of a Generic Model of Eukaryotic Cell-Cycle Regulation". *Biophysical Journal* 90 (2006), pp. 4361–4379. 10.1529/biophysj.106.081240.

[36] K. C. Chen *et al.* "Kinetic Analysis of a Molecular Model of the Budding Yeast Cell Cycle". *Molecular Biology of the Cell* 11 (2000). Ed. by M. J. Solomon, pp. 369–391. 10.1091/mbc.11.1.369.

[37] J. D. Meiss. *Differential dynamical systems*. Vol. 14. Siam, 2007.

[38] J. D. Davis, C. M. Kumbale, Q. Zhang, and E. O. Voit. "Dynamical systems approaches to personalized medicine". *Current Opinion in Biotechnology* 58 (2019), pp. 168–174. 10.1016/j.copbio.2019.03.005.

[39] M. Barberis *et al.* "Sic1 plays a role in timing and oscillatory behaviour of B-type cyclins". *Biotechnology Advances* 30 (2012), pp. 108–130. 10.1016/j.biotechadv.2011.09.004.

[40] C. Linke *et al.* "A Clb/Cdk1-mediated regulation of Fkh2 synchronizes CLB expression in the budding yeast cell cycle". *npj Systems Biology and Applications* 3 (2017), p. 7. 10.1038/s41540-017-0008-1.

[41] M. A. Savageau *et al.* "Phenotypes and tolerances in the design space of biochemical systems." *Proceedings of the National Academy of Sciences of the United States of America* 106 (2009), pp. 6435–40. 10.1073/pnas.0809869106.

[42] B. A. Benayoun, S. Caburet, and R. A. Veitia. "Forkhead transcription factors: Key players in health and disease". *Trends in Genetics* 27 (2011), pp. 224–232. 10.1016/j.tig.2011.03.003.

[43] K. D. MacIsaac *et al.* "An improved map of conserved regulatory sites for Saccharomyces cerevisiae." *BMC bioinformatics* 7 (2006), p. 113. 10.1186/1471-2105-7-113.

[44] B. J. Venters *et al.* "A Comprehensive Genomic Binding Map of Gene and Chromatin Regulatory Proteins in Saccharomyces". *Molecular Cell* 41 (2011), pp. 480–492. 10.1016/j.molcel.2011.01.015.

[45] A. Z. Ostrow *et al.* "Fkh1 and Fkh2 Bind Multiple Chromosomal Elements in the S. cerevisiae Genome with Distinct Specificities and Cell Cycle Dynamics". *PLoS ONE* 9 (2014). Ed. by Y. Wang, e87647. 10.1371/journal.pone.0087647.

[46] H. S. Rhee and B. F. Pugh. "Comprehensive Genome-wide Protein-DNA Interactions Detected at Single-Nucleotide Resolution". *Cell* 147 (2011), pp. 1408–1419. 10.1016/j.cell.2011.11.013.

[47] A. L. Lehninger. *Bioenergetics: The Molecular Basis of Biological Energy Transformation*. WA Benjamin, 1965.

[48] R. van der Meer, H. Westerhoff, and K. Van Dam. "Linear relation between rate and thermodynamic force in enzyme-catalyzed reactions". *Biochimica et Biophysica Acta (BBA) - Bioenergetics* 591 (1980), pp. 488–493. 10.1016/0005-2728(80)90179-6.

[49] K. Schuchmann and V. Müller. "Autotrophy at the thermodynamic limit of life: a model for energy conservation in acetogenic bacteria". *Nature Reviews Microbiology* 12 (2014), pp. 809–821. 10.1038/nrmicro3365.

[50] N. C. Duarte *et al.* "Global reconstruction of the human metabolic network based on genomic and bibliomic data". *Proceedings of the National Academy of Sciences* 104 (2007), pp. 1777–1782. 10.1073/pnas.0610772104.

[51] I. Thiele *et al.* "A community-driven global reconstruction of human metabolism". *Nature Biotechnology* 31 (2013), pp. 419–425. 10.1038/nbt.2488.

[52] T. Shlomi, M. N. Cabili, and E. Ruppin. "Predicting metabolic biomarkers of human inborn errors of metabolism". *Molecular Systems Biology* 5 (2009), p. 263. 10.1038/msb.2009.22.

[53] O. Warburg. "On the Origin of Cancer Cells". *Science* 123 (1956), pp. 309–314. 10.1126/science.123.3191.309.

[54] E. Brunk *et al.* "Recon3D enables a three-dimensional view of gene variation in human metabolism". *Nature Biotechnology* 36 (2018), pp. 272–281. 10.1038/nbt.4072.

CHAPTER 2

Clb3-centered regulations are recurrent across distinct parameter regions in minimal autonomous cell cycle oscillator designs

2.1 Introduction **18**
2.2 Results **22**
Experimental rationale underlying the computational analyses 22
The minimal cell cycle model and derivation of designs 1A–3 . 23
Conserved network motifs across oscillatory phenotypes 25
Alternative network designs of the minimal cell cycle model . . 27
The ability of extended designs to generate oscillations 30
Limit cycles belong to distinct parameter space regions 31
2.3 Discussion **36**
2.4 Methods **41**
Simulation of Ordinary Differential Equation (ODE) models . . 42
Application of the SDS Toolbox 42
Parameter space sampling to find oscillatory phenotypes 42
Principal component analysis 44
Supplementary Information **46**

Adapted from:
Mondeel, T.D.G.A., Ivanov, O., Westerhoff, H.V., Liebermeister W., Barberis M. Clb3-centered regulations are recurrent across distinct parameter regions in minimal autonomous cell cycle oscillator designs. npj Syst Biol Appl 6, 8 (2020). 10.1038/s41540-020-0125-0.

> "Rather grandly I argued to myself that the process of reproduction was a central property of life, and that this was seen in its simplest form with the reproduction of cells. Therefore, I reasoned that study of the cell cycle responsible for the reproduction of cells was important and might even be illuminating about the nature of life."
>
> — Paul Nurse[1]

Abstract

Some biological networks exhibit oscillations in their components to convert stimuli to time-dependent responses. The eukaryotic cell cycle is such a network, being governed by waves of cyclin-dependent kinase (cyclin/Cdk) activities that rise and fall with specific timing and guarantee its timely occurrence. Disruption of cyclin/Cdk oscillations could result in dysfunction through reduced cell division. Therefore, it is of interest to capture the properties of network designs that exhibit robust oscillations. Here we show that a minimal cell cycle network is able to oscillate autonomously, and that cyclin/Cdk-mediated positive feedback loops (PFLs) and Clb3-centered regulation is recurrent for sustained cyclin/Cdk oscillations in 11 known and hypothetical network designs. We propose that Clb3-mediated coordination of cyclin/Cdk waves reconciles checkpoint and oscillatory cell cycle models. Considering the evolutionary conservation of the cyclin/Cdk network across eukaryotes, we hypothesize that functional ('healthy') phenotypes require the capacity to oscillate autonomously compared to dysfunctional (potentially 'diseased') phenotypes.

2.1 Introduction

Living systems exhibit dynamic self-organization, i.e. the spontaneous emergence of spatio-temporal order with the formation of various spatio-temporal patterns [1]. Self-organization may involve oscillations in the concentrations of a system's components [2–4], which have been observed at various temporal scales. Oscillatory behavior arises from non-linear interactions among two or more components of a system [5]. An example is given by the eukaryotic cell cycle, the sequential process through which a growing cell replicates and divides into two daughter cells. The dynamics of this process are implemented through biochemical interactions between genes and proteins, and are governed by periodic waves of cyclin-dependent kinase (Cdk) activities [6–10].

Here, self-organization in the form of oscillations results from the sequential activation and inactivation of a number of cyclin/Cdk complexes that regulate a timely cell cycle [8]. The periodic fluctuations of cyclin/Cdk activities are regulated by cyclin levels (i) through transcription factors and (ii) through targeted

[1] https://www.nobelprize.org/prizes/medicine/2001/nurse/biographical/

degradation by multi-protein complexes such as the Anaphase-Promoting Complex (APC).

Sustained cyclin/Cdk oscillations equate to growth and cell division. For bacteria and single cell organisms such as budding yeast, (faster) growing subpopulations will outperform slower growing and not-growing subpopulations, thus providing a selectable advantage. Thus, the increased fitness for an organism, realized through sustained, autonomous oscillations, can be considered a functional or 'healthy' phenotype of a cell. In contrast, lack of oscillations in cyclin/Cdk complexes is to be considered dysfunctional or "diseased" behavior, unless quiescent cells are considered.

Mathematical modeling can be of help to better understand how cell cycle networks exhibit oscillations with certain properties, e.g. a specific amplitude and/or frequency and a definite order of appearance among a system's components, mathematical modeling may be performed. Cell cycle oscillations may be modeled (i) by sustained oscillations in the form of limit cycles, where cyclin/Cdk oscillations arise independently from external factors, or (ii) by checkpoint mechanisms, where external requirements such as attaining a minimum cell size to progress from G1 to S phase are explicitly taken into account in the form of irreversible transitions between steady-states. Here, checkpoints act as signals that delay the cell cycle phase transitions by stabilizing the dynamics in alternative stable steady-states of the underlying biochemical system. Contrarily, sustained autonomous oscillations exhibit limit cycles around a single steady-state. The checkpoint view is currently prevalent, due to correlations observed between the cell cycle period and the growth rate [11], although noise-induced oscillations have been theoretically predicted when cell size is constant [12]. However, models that exhibit autonomous oscillations in the form of limit cycles are better suited when networks are investigated in absence of external controls such as cell size.

Among the network designs that have been described to characterize cell cycle oscillators, positive feedback loops (PFLs) enhance amplitude and robustness of cyclin/Cdk [13–15]. PFLs promote switch-like responses that guarantee unidirectionality of cell cycle progression [5]. Similarly, negative feedback loops (NFLs) are considered necessary for oscillations to occur [15, 16], and a model consisting of at least three ordinary differential equations is needed for sustained oscillations to occur [5]. It has been conjectured that PFLs have evolved to facilitate oscillations in NFLs at lower, kinetically achievable, degrees of cooperativity [5]. Alteration in the frequency of cyclin/Cdk oscillations or of a cell cycle as a whole may correspond to alteration of cell proliferation, thereby to a dysfunctional or 'disease' phenotype of a cell, as a result of deregulation of timely cyclin/Cdk activities [9, 17, 18]. This deregulation may impinge on the cellular concentrations of cyclin and Cdk proteins, which already exhibit significant oscillations in a wild type cell [19].

Here we build on our previously published minimal model of the cell cycle network [10] to generate the first truly autonomously oscillating model of Clb/Cdk1 complexes in budding yeast, with the intent to: (i) simplify our previously published model to make it more amenable to the parameter scans per-

formed in this work, (ii) integrate new evidence in order for the model to accurately reflect the experimental observations, and (iii) investigate the effect of hypothetical interactions that can be validated experimentally. For each of the 11 resulting network designs we investigate (i) which network designs exhibit autonomous, stable oscillations, i.e. limit cycles, and (ii) how network designs and associated parameters influence the occurrence of these oscillations.

The design described by our model's comprehends: (i) three cyclin/Cdk complexes, i.e. Clb5,6/Cdk1, Clb3,4/Cdk1 and Clb1,2/Cdk1, which exert their function in the S-G2-M (mitotic) phases of the cell cycle, and (ii) their stoichiometric inhibitor Sic1 that is active in G1 phase. One key feature of our model design is the incorporation of the Clb3 cyclin, which is lacking in existing cell cycle models [20, 21]. Our analysis is driven by the hypothesis that elements of the interaction network are critical, or more important, than others to generate sustained oscillations. Dynamic models based on this design exhibit transient oscillations of all mitotic cyclins simultaneously [10], and thereby of cyclin/Cdk activities, which may result in a frequency characteristic of a functional wild type cell. However, so far no analysis has been conducted to investigate: (i) whether this specific model is able to oscillate autonomously, and (ii) how the occurence and properties of cyclin/Cdk oscillations can be modulated by variations of model parameters, suggesting shifts to dysfunctional or hyperfunctional states.

Other models of the cell cycle have been investigated in terms of their potential to show oscillations (briefly called "oscillatory potential" below) [8, 20, 21]; however, these are: (i) larger in size, (ii) different in the network structure, and (iii) analyzed only using conventional bifurcation analysis techniques to find single oscillating points or regions in the parameter space. Due to the differences in network design and model size, it is not clear a priori that the results from existing models would translate to the simplified network investigated here.

Several methods exist for finding parameter sets leading to bifurcations and oscillations in biochemical networks [22–24]. In this work, we make use of the System Design Space (SDS) methodology [25–27] to detect limit cycles more easily [28], and analyze the ability of our minimal cell cycle model to generate transient and sustained oscillations. The application of the SDS methodology to analyze oscillatory behavior is novel in the cell cycle field, and it has never been applied to models of the size considered here. Identifying limit cycles is an open mathematical challenge for high-dimensional systems, and even numerically this is challenging. This problem is exacerbated if the interest is to find multiple limit cycles across distinct regions in the parameter space, as the existing methods [22–24] do not accommodate this aspect as easily as the SDS. Our pipeline centered around the SDS method allows to search for oscillations across a set of regions, each with unique network properties, that partition the parameter space. The SDS methodology relates genotype and environment, which affect biochemical and environmental parameters in the system, to the phenotype of steady-state attributes of the biochemical system, by deconstructing the biochemical system into a finite number of qualitatively distinct subsystems. Within this approach, the term 'phenotype' refers to a combination of 'dominant terms', i.e. a subset of interactions in the network that are large in numerical value with respect to the

other terms, which are neglected from the equations. Note that the parameters (genotype and environment) and dynamic concentrations in the system define which terms are dominant (numerically large) and therefore which phenotype is expressed.

The computational cost of using the SDS methodology increases with the number of terms in the model equations, since this translates into more distinct phenotypes. For this reason, it is advantageous to use our previously published model, which has significantly less terms than other published models for budding yeast [20, 21, 29]. Applying the pipeline considered here to the more complex yeast cell cycle models would most likely require significant computer cluster usage. The disadvantage of using a minimal model is that biochemical details that have been uncovered about the cell cycle regulatory network may be lacking. However, one may expect that if the core design of a minimal and detailed models are similar, the general properties remain the same as well as shown for both budding yeast [29] and fission yeast [30]. Furthermore, the implementation time of the complex, yet powerful, framework provided by the SDS methodology is greatly simplified by utilizing the Systems Design Space Toolbox [31].

In this work, we present the first autonomously oscillating Clb/Cdk1 model for budding yeast, and have explored the oscillatory behavior of 11 known and hypothetical network designs. We recovered the known importance of PFLs and NFLs for oscillations. More specifically, we show that a positive feedback loop (PFL) by Clb3/Cdk1 on *CLB3* synthesis (Clb3 PFL) improves the ability of our models to produce sustained Clb/Cdk1 oscillations, and that a positive feedback loop by Clb2/Cdk1 on CLB2 synthesis (Clb2 PFL) takes over this key role when the model takes into account the inhibition of G1/S cyclins by Clb2/Cdk1. Furthermore, we show that two regulatory activations, i.e. Clb5 $\rightarrow$ Clb3 and Clb3 $\rightarrow$ Clb2, forming a transcription factor-mediated linear *CLB* cascade that we have recently discovered [32], are more frequently dominant in phenotypes that yield sustained Clb/Cdk1 oscillations as compared to the feed-forward Clb5 $\rightarrow$ Clb2 regulation described earlier [33]. We thus hypothesize that functional ("healthy") phenotypes require the capacity to oscillate autonomously – through Clb3-centered regulations – compared to dysfunctional (potentially 'diseased') cellular phenotypes - where these designs are altered and the potential for oscillatory behavior is reduced. We envision a scenario in which Clb5 and Clb2 are involved in the checkpoints, whereas Clb3-centered regulations that coordinate Clb5 and Clb2 drive autonomous cell cycle oscillations to maintain cell proliferation. This scenario thus reconciles checkpoint and oscillatory views of cell cycle regulation. In addition, we highlight that the transcriptional inhibition of G1/S cyclins and Sic1 by mitotic Clb/Cdk1 results in particularly strong NFLs for stabilizing oscillations. Finally, through perturbation of selected limit cycles, we identify crucial model parameters that exert the strongest control on the frequency of Clb/Cdk1 oscillations.

Given the evolutionary conservation of the cell cycle network across eukaryotes, the mitotic cyclin/Cdk network can be used as a core building block of multi-scale models that integrate regulatory modules to address cellular physiology.

2.2 Results

Experimental rationale underlying the computational analyses

The cell cycle has a unique property as compared to other biochemical networks. Its drivers, i.e. the cyclin subunits that regulate the Cdk activity, have both specialized functions as well as partially overlapping functions, through different specificity of binding to the substrates that they recognize and – through their partner Cdk – phosphorylate [34]. Budding yeast cells lacking Clb5 (S phase cyclin) do not replicate at the proper time, but they do so progressively after activation of Clb2 (G2/M phase cyclin), which can partially substitute for the missing Clb5 activity; this indicates that a partial overlap in the cyclin function helps to drive DNA replication [35]. In these cells, S phase is prolonged and the overall cell cycle timing is slightly delayed [36]. Conversely, cells lacking Clb2 (G2/M phase cyclin) exhibit defects in mitotic entry and delay in mitotic exit [37]; moreover, modified Clb2 degradation kinetics result in a compromised viability [38]. In these cells, Clb5 (S phase cyclin) and/or Clb3 (S/G2 cyclin) cannot substitute for the missing Clb2 activity, indicating the relevance of cyclin specificity for the events that trigger cell division.

Differently from Clb5 and Clb2, cells lacking Clb3 or cells where Clb3 degradation kinetics have been modulated are viable and complete cell division at the same timing as a wild type cell [38]. In fact, Clb2 can replace Clb3 activity (Clb2 replaces Clb3 better than it does with Clb5, as Clb2 and Clb3 have more structural and functional similarities than Clb2 and Clb5). Whereas Clb5 and Clb2 deletions affect dynamics of cell division timing as well as cell viability, Clb3 deletion does not affect cell cycle timing nor cell viability. Clb3 deletion is lethal only in the clb2Δ clb3Δ double mutant [37], and in the clb5Δ clb3Δ clb4Δ [39] and clb2Δ clb3Δ clb4Δ [37, 40–42] triple mutants, indicating that Clb5 and Clb2, respectively, are required for spindle formation in the absence of Clb3 and Clb4.

Taking into account this experimental evidence, we envision a scenario where (i) Clb5 and Clb2 serve a function in checkpoint models (as currently incorporated in Tyson/Novák's cell cycle models [20, 21]), whereas (ii) Clb3 serves a function in autonomous oscillations required to sustain the cell's viability. Specifically: (i) In Tyson/Novák's cell cycle models, Clb5 and Clb2 represent the checkpoints that drive the cell cycle through the next cell cycle phase, should their concentration reach a definite threshold. In the cell, DNA damage/errors would activate the checkpoint affecting Clb5 levels, thus slowing/halting DNA replication dynamics, whereas troubles in cell division would activate the checkpoint affecting Clb2 levels, thus delaying/impairing cell division. In addition, the requirement of definite Clb5/Clb2 threshold concentrations may be seen as the result of a proper availability of nutrients which, if lacking, would not allow the thresholds to be reached, thus the cell cycle not to be completed. Conversely: (ii) Clb3 has never been considered in any existing (checkpoint) models of cell cycle regulation, possibly due to its not fully clear and not critical role in cell division. In our view, Clb3 serves a function in the cell's autonomous oscillations. Clb3 is not involved in the checkpoints, as its deletion is lethal only in the clb5Δ

clb3Δ clb4Δ [39] and clb2Δ clb3Δ clb4Δ [40–42] triple mutants, but not in the clb5Δ clb3Δ and clb2Δ clb3Δ [40] double mutants. Furthermore, we have discovered, through a detailed computational and experimental investigation, the role of Clb3 in the coordination of the mitotic waves of cyclins, synchronized from the S through M phases in a linear cascade (Clb5 → Clb3 → Clb2) through the Fkh2 transcription factor [32]. In order to shed light on this hypothesis, in this study we have conducted a detailed computational analysis, to investigate the occurrence of sustained oscillations in a minimal Clb/Cdk1 model and to identify recurring Clb-mediated principles of design, i.e. network motifs, underlying autonomous oscillations.

The minimal cell cycle model and derivation of designs 1A–3

Starting from our previously published minimal cell cycle model [10] (*Design 1A*, Supplementary Information, Fig. S2.1A), we built a number of mathematical models based on Ordinary Differential Equations (ODEs) that are able to generate oscillations. The core design considers four species: (i) three representing the complexes that Cdk1 forms with the three pairs of B-type cyclins, Clb5,6, Clb3,4 and Clb1,2, and (ii) the inhibitor Sic1 that binds and inhibits all three Clb/Cdk1 complexes. Each of the four species is associated to cell cycle events during a specific phase of the cell cycle (Fig. 2.1A). The model describes (i) the progressive activation of the three Clb/Cdk1 complexes in a linear cascade, and (ii) the complex formation between Sic1 and the Clb/Cdk1 complexes, which are mutually inhibiting one another (Fig. 2.1B). The minimal model considers the complexity of all documented interactions among the Clb/Cdk1 complexes (Table 2.2 and Fig. 2.1C, solid lines) in addition to two hypothetical interactions (Table S2.2 and Fig. 2.1C, dotted lines). In addition, the model describes the degradation of the Clb/Cdk1/Sic1 complex, and the basal synthesis and degradation of each species, as visualized in the interaction diagram (Fig. 2.1D).

The existing model [10] is rooted in experimental evidence, and the new models presented here were built considering: (i) recently unraveled experimental evidence, (ii) hypotheses generated on existing experimental evidence, and (iii) simplification of a number of reactions (Table S2.2). This process resulted in five sequential model designs: 1A, 1B, 1C, 2 and 3. The essential differences between the five designs are summarized in Fig. 2.1D and S2.1). In Supplementary Information, Section 2.4, we document the step-by-step derivation of the five alternative network designs. *Design 3* presents a special case as it incorporates a novel quasi-steady-state approximation, which assumes that formation (k^+) and/or dissociation (k^-) of the Clb/Cdk1/Sic1 ternary complexes occur on a faster time-scale than the other processes considered in the model (Mart Loog, personal communication, and Supplementary Information, Section 2.4.

The main aim of this work is to systematically identify limit cycles across distinct parameter regions and across multiple minimal model designs, in order to identify network motifs (presence of specific interactions and parameter values) that support the occurrence of oscillations. As a preliminary analysis, we investigated the ability of designs 1A–3 to generate: (i) transient cyclin/Cdk oscillations

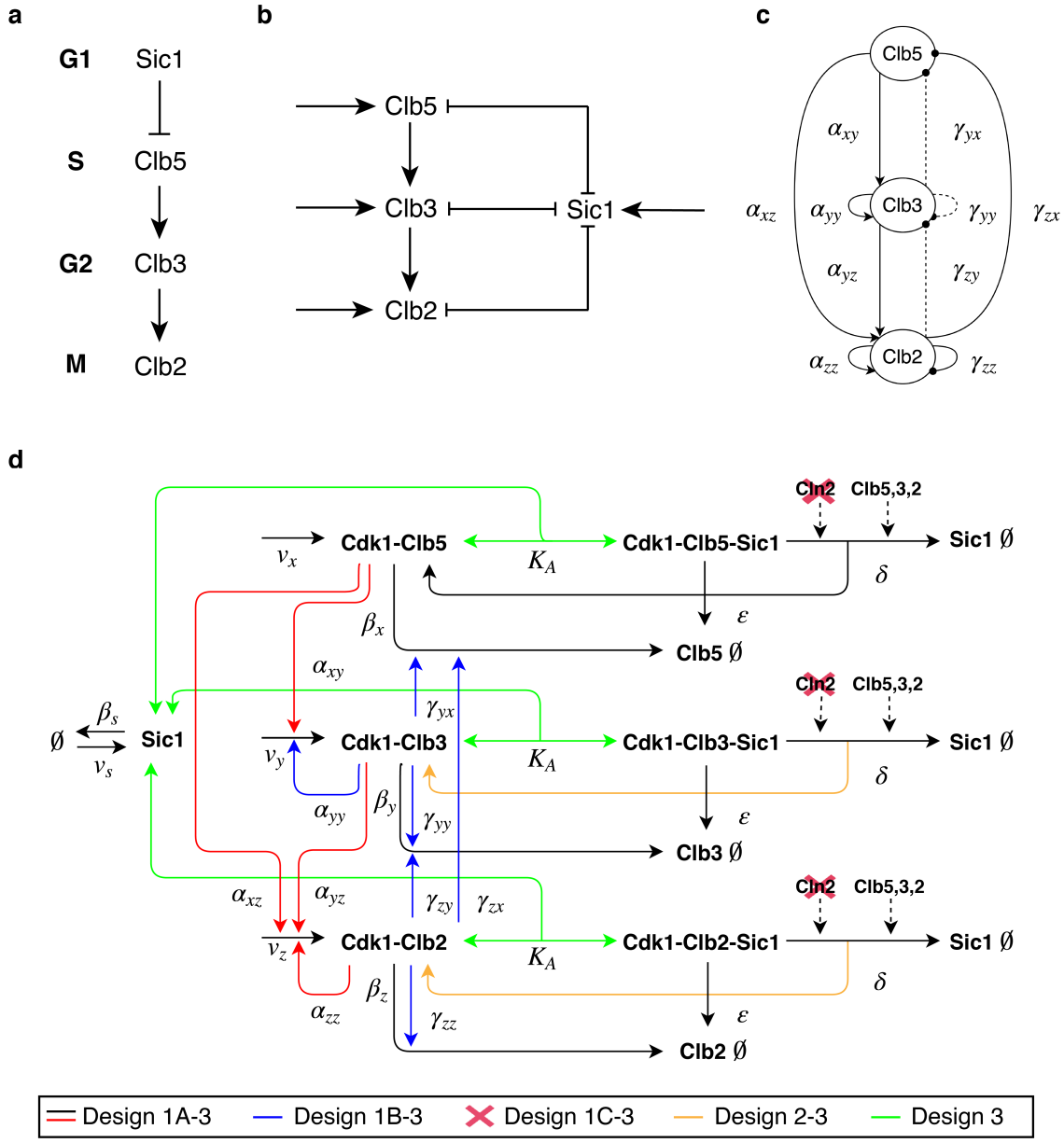


Figure 2.1: Schematic views of the minimal cell cycle model for budding yeast and full interaction diagram for designs 1A, 1B, 1C, 2 and 3. (**a**) Key molecular players driving phase-specific cell cycle events. (**b**) Linear cascade between the three Clb/Cdk1 complexes, and mutual inhibition between these and the Clb/Cdk1 inhibitor Sic1. (**c**) Interactions among the Clb/Cdk1 complexes. Solid lines indicate (8) proven interactions, whereas dotted lines indicate (2) hypothetical interactions (see Table S2.2). (**d**) Full interaction diagram for designs 1A, 1B, 1C, 2 and 3 of the minimal cell cycle network. The scheme illustrates the core interactions in all model designs presented in this work, i.e. black and red arrows for the basal and activatory regulations, respectively, and highlights the progressive changes to the core structure introduced in designs 1A–3 (blue, red cross, orange and green, respectively). Dotted arrows indicate the Cln(/Cdk1)- and Clb(/Cdk1)-mediated phosphorylation of Sic1 in Clb/Cdk1/Sic1 ternary

Figure 2.1: (Continued) complexes, resulting in its degradation. The complex formation between Clb/Cdk1 complexes and Sic1 is indicated with the K_A parameter, referring to the quasi-steady-state assumption introduced in *Design 3* (see Supplementary Information, Section 2.4, which should be taken to be the regular complex formation (k^+ for formation, k^- for dissociation) for designs *1A–2*. Model derivations are reported in Supplementary Information, Section 2.4, and details of the reactions and their experimental evidence are reported in Table S2.2.

and (ii) sustained oscillations in the form of limit cycles. Finally, for each design, we performed a sensitivity analysis of how the model parameters influence the period of a single limit cycle. We observed that all five model designs are able to generate limit cycles, and that the basal synthesis and degradation parameters together with those responsible for the linear Clb cascade (Clb5 -> Clb3 -> Clb2), hold the greatest control over the length of the period of oscillation (see Supplementary Information, Section 2.4).

Conserved network motifs across oscillatory phenotypes

For a system of the size of model designs *1A-3* (7 species and over 20 parameters), there is generally no way to obtain all possible parameter sets that give rise to limit cycles, and an infinite number of these could exist. However, a relevant question, interpretable biologically, is to address which network designs are able to generate oscillatory behavior in the form of limit cycles. The System Design Space (SDS) methodology [25] can be utilized to investigate such a capability of definite network designs. Specifically, the methodology allows to partition network designs into 'phenotypes' that represent the dominance of certain reactions over others, neglecting all non-dominant ones. For a given set of parameters and concentrations (i.e. states) one activatory term and one inhibitory term in each differential equation are numerically larger than all others, i.e. dominant (see the Methods section 2.4 and Supplementary Information, Section 2.4). The dominance of a reaction implies boundary conditions, i.e. inequalities in the parameter and state spaces, which, if feasible, partition these spaces into areas referred to as 'valid' phenotypes. Parameter sets that yield oscillations occur within a valid phenotype and therefore link the dominance of specific reactions to the occurrence of oscillations.

In order to analyze models for the occurrence of limit cycle oscillations, we implemented a modeling pipeline based on the SDS methodology that incorporates a novel approach to sample phenotypes for finding limit cycles, applicable to any network design. We implemented a parameter sampling procedure that makes use of the boundaries of a phenotype and employs log-uniform random sampling (see Fig. 2.2 and the Methods section 2.4).

We set up model definitions in the GMA form (see Supplementary Information, Section 2.4) for all the kinetic models representing five network designs considered in this work. In addition, we used the model presented by Savageau and colleagues [43] as a test case to make sure that our implementation

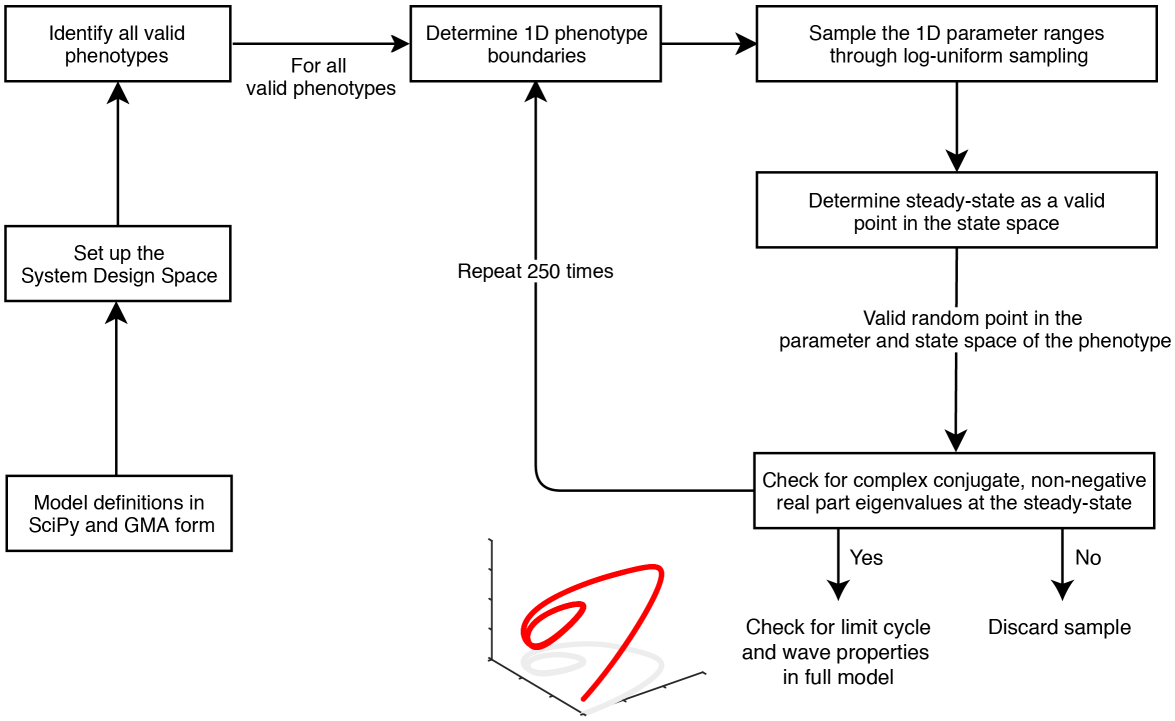


Figure 2.2: Computational pipeline implemented by using the System Design Space Toolbox, to identify oscillatory phenotypes in high-dimensional parameter spaces, for any model, using targeted parameter sampling. The procedure starts from a previously defined set of model equations in the format used by Scipy and the System Design Space Toolbox. All valid phenotypes with consistent boundary conditions in each model were identified. Each such valid phenotype was sampled log-uniformly 250 times. For each random sample, the parameter values were checked to lie within the parameter region defined by the phenotype and the steady-state was calculated. The potential for oscillation was identified by looking for the presence of two non-negative eigenvalues in the phenotype steady-state. If this condition was met, the full model was analyzed for the presence of a limit cycle (see "Methods" section 2.4).

could recover their previously reported results (data not shown). By using our pipeline (Fig. 2.2) we generated the System Design Space for each model variant, i.e. the set of all phenotypes, and retrieved the set of valid phenotypes (see Supplementary Information, Section 2.4). For our models, the valid phenotypes represented $0.044\% - 0.34\%$ of all theoretically possible phenotypes. This reduced the number of phenotypes for which the stability was investigated to a manageable amount of several hundred to several thousand phenotypes (Table 2.1). After identifying the valid phenotypes, these were sampled to retrieve parameter sets yielding potential oscillations by using log-uniform random sampling. As a criterion for oscillations that emerge from a Hopf bifurcation, for each sampled parameter set we checked for the presence of a pair of complex conjugate eigenvalues with non-negative real part in the steady-state of a phenotype. If a particular combination of phenotype and parameter set satisfied this condi-

tion, and therefore showed potential for oscillations, the limit cycle behavior in the full kinetic model was then tested using that specific parameter set.

We observed that in *Design 3*, with 250 samples for each valid phenotype, 664 phenotypes with a potential for oscillations were identified, and 8 limit cycles across 7 of these were found. Conversely, for *Design 1A* no positive complex conjugate eigenvalues were found, supporting the existing point of view that NFLs are required for sustained oscillations [15, 16]. Designs *1B-2* exhibit limit cycles but with an incorrect order of peaking of the four model species so we did not count them. The results for the updated model *Design 3* can be directly compared to designs *1A* and *1B*, which are based on our published minimal model [10, 32]; we conclude that the updated model outperforms its counterparts.

Each parameter set that yielded a limit cycle was stored, and the time-dependent oscillatory behavior was plotted (the parameter sets are available in the Supplementary Code Repository). For each phenotype for which a limit cycle was found, the terms in the differential equations that were dominant for that phenotype were identified. Inspection of these dominant processes allowed for counting the existence of specific parameters within these phenotypes (Table 2.2). In *Design 3*, α_{yy} (Clb3 PFL, responsible for Clb3/Cdk1 activation) and α_{yz} (Clb2/Cdk1 activation by Clb3/Cdk1) are the activatory parameters observed most frequently in phenotypes that yielded limit cycles (Table 2.2). This finding suggests that the dominance of these two terms in the differential equations increases the ability to generate sustained oscillations, perhaps by enlarging the region within phenotypes of the design space where oscillations occur. The result confirms the relevance of the linear *CLB* cascade through the Fkh2 transcription factor that we have recently discovered [32] - formed by the two regulatory activations Clb5 $\rightarrow$ Clb3 (α_{xy}) and Clb3 $\rightarrow$ Clb2 (α_{yz}) - over the Clb5 $\rightarrow$ Clb2 regulation (α_{xz}) described earlier [33]. Altogether, these findings point to the relevance of Clb3 for generating sustained Clb/Cdk1 oscillations, through the dominance (i) of the Clb3 PFL and (ii) of the linear cascade (Clb5 $\rightarrow$ Clb3 $\rightarrow$ Clb2).

With respect to the inhibitory regulations, we observed that the Clb3 NFL (γ_{yy}) and the Clb2 NFL (γ_{zz}) terms are rarely dominant in phenotypes exhibiting limit cycles (Table 2.2). Conversely, the parameters referring to the APC-mediated inhibition of Clb5 and Clb3 by Clb2 (γ_{zx} and γ_{zy}, respectively), and to the degradation rates of Sic1 and Clbs from the Clb/Cdk1/Sic1 ternary complex (δ and ϵ, respectively) were observed more frequently with respect to the generation of sustained cyclin/Cdk1 oscillations.

Alternative network designs of the minimal cell cycle model

To further analyze the oscillatory behavior of our minimal cell cycle model, we extended *Design 3* to test six new network designs that include further known or hypothetical inhibitory regulations. By doing this, we aim to understand whether and how these regulations might enable the cyclin/Cdk network to generate limit cycles. The new designs are based on *Design 3*, and are referred to as *Design 4* through *Design 9* (Fig. 2.3). Each new design reflects either a single or several related inhibitory regulations. *Design 4*, *Design 5* and *Design 6* describe known in-

Design #	Total phenotypes	Valid phenotypes	Phenotypes with oscillatory potential	Phenotypes with limit cycles	Total limit cycles
3	995328	3355	664	7	8
4	1990656	6689	1158	66	204
5	1990656	6950	1523	16	23
6	1990656	6844	1508	20	31
7	1990656	6692	1239	19	24
8	1990656	6750	974	9	9
9	1990656	6977	677	13	21

Table 2.1: Number of phenotypes: in total, that are valid, that show potential for oscillations (presence of two non-zero eigenvalues). The number of limit cycles retrieved and the number of distinct phenotypes these occured in for model *Design 3* through *Design 9*.

Design #	α_{xy}	α_{yy}	α_{xz}	α_{yz}	α_{zz}	γ_{yx}	γ_{zx}	γ_{yy}	γ_{zy}	γ_{zz}	K	δ	ϵ	β_x	β_y	β_z	β_s	v_y	v_z
3	0	6	0	6	1	1	5	1	4	0	–	7	3	1	0	4	0	1	0
4	41	23	1	17	48	12	20	12	18	4	54	46	63	14	11	5	0	2	0
5	3	12	2	10	4	0	11	0	13	0	8	14	12	5	1	4	0	1	0
6	3	17	0	13	7	3	13	2	12	1	10	16	15	3	2	4	2	0	0
7	6	12	2	9	7	3	14	3	12	0	16	18	13	2	1		0	1	1
8	2	6	3	3	2	2	6	0	6	2	2	9	6	1	0	3	0	1	1
9	2	11	1	9	3	2	11	1	10	1	1	12	7	0	1	5	0	0	0

Table 2.2: Counts of the occurrence of parameters in dominant terms of phenotypes that yielded limit cycles for model *Design 3* through *Design 9*. The counts listed are subsets of the numbers in the column "Phenotypes with limit cycles" in Table 2.1. Parameters that are present in all phenotypes, due to the model design, are not shown (v_s, v_x, k_A). The generic K parameter modulates the strength of the unique novel inhibitions in designs *4–9*. For example, in *Design 4*, K represents the parameter K_{zx}, which refers to the transcriptional inhibition of Clb2/Cdk1 (z) on Clb5/Cdk1 (x). Parameters that are part of terms that were dominant in more than 60% of all phenotypes that yielded limit cycles per design (Table 2.1), are highlighted and color coded in groups: activatory (red) and inhibitory (blue) interactions among the Clb/Cdk1 complexes, novel inhibitory interactions in designs *4–9* (yellow), and basal synthesis and degradation reactions (green).

hibitory regulations mediated by Clb/Cdk1 activities (Fig. 2.3A), whereas *Design 7*, *Design 8* and *Design 9* describe hypothetical inhibitory regulations mediated by Sic1 (Fig. 2.3B). In the following, each design is described succinctly, and both detailed molecular mechanisms and equations supporting the designs are reported in Supplementary Information, Section 2.4.

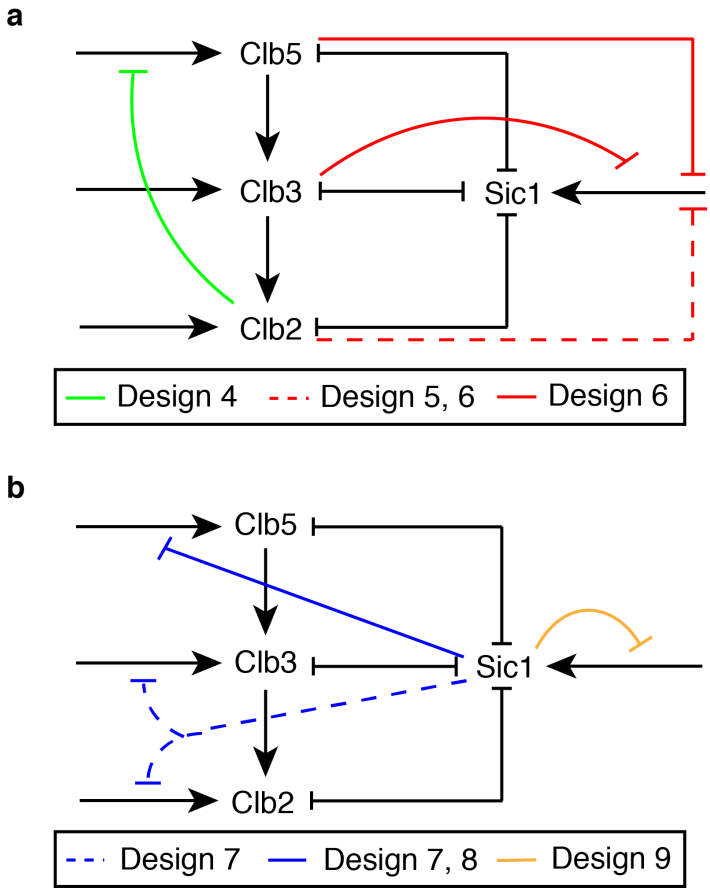


Figure 2.3: Schematic view of known and hypothetical inhibitory regulations added to *Design 3* of the minimal cell cycle model. (**a**) *Design 4*, *Design 5* and *Design 6* describe known inhibitory regulations mediated by Clb/Cdk1 activities. (**b**) *Design 7*, *Design 8* and *Design 9* describe hypothetical inhibitory regulations mediated by Sic1. Colored lines indicate novel regulations, with each color identifying a particular network design. Dashed lines indicate regulations occurring in two different designs. In the latter, the two designs related to each such interaction are shown with the same color. Black lines indicate the activatory and inhibitory regulations occurring in the minimal cell cycle model. See text and Supplementary Information, Section 2.4 for details about the molecular mechanisms.

Design 4 incorporates the inhibition of Clb5/Cdk1 by Clb2/Cdk1 through the MBF transcription factor, formed by Mbp1 and Swi6. Clb2 has been shown to interact physically with Swi4, and to repress transcription of the G1 cyclins [44]. This inhibition translates to an effective inhibition of the Clb5/Cdk1 activity, due to the lack of the PFL between the G1 phase Cln2/Cdk1 complex and SBF/MBF

[45] and to the lifted inhibition of Sic1 by Cln1,2/Cdk1 [46]. *Design 5* and *Design 6* incorporate the inhibition of Clb/Cdk1 on *SIC1* transcription through the *SWI5* transcription factor. Specifically, *Design 5* describes the inhibition of *SIC1* transcription mediated by the Clb2/Cdk1 activity [47], reflecting the likely scenario where the most abundant Cdk1 activity is due to Clb2/Cdk1. *Design 6* describes the same mechanism mediated by the three Clb/Cdk1 complexes: Clb2/Cdk1, Clb3/Cdk1 and Clb5/Cdk1.

Design 7 and *Design 8* incorporate the hypothetical inhibition of Sic1 on mitotic *CLB* transcription to rationalize a recent observation that Sic1 oscillations rescue viability of cells with low levels of mitotic Clb cyclins [48]. Specifically, *Design 7* describes the inhibition of Clb2 and Clb3 synthesis - which we have recently shown to be regulated by a similar transcriptional mechanism [32] - by Sic1. *Design 8* describes the hypothetical inhibition of Clb5, Clb3 and Clb2 syntheses by Sic1. Finally, *Design 9* incorporates the hypothetical inhibition of Sic1 synthesis by a Sic1-mediated NFL through *SWI5*.

The ability of extended designs to generate sustained oscillations

To retrieve limit cycles for the new network designs *4-9*, we again employed the pipeline described in Fig. 2.2. Each design yielded several hundred phenotypes that corresponded to parameter space regions with two non-negative complex conjugate eigenvalues, and all designs yielded a set of limit cycles (Table 2.1). Specifically, all designs yielded a higher number of limit cycles and more distinct phenotypes with limit cycles than *Design 3* did. This points to a stronger tendency to oscillate due to the new inhibitory regulations. Interestingly, *Design 4* outperformed all other designs in terms of the number of limit cycles retrieved, followed by *Design 6*, *Design 7* and *Design 5*. The computational results obtained for *Design 4* and *Design 5* indicate a role in generating and stabilizing sustained oscillations for the two inhibitory regulations experimentally observed: (i) Clb2/Cdk1 on Clb5/Cdk1, indirectly, through Swi4 [44], and (ii) Clb2/Cdk1 on Sic1, directly, through Swi5 [47].

Among the hypothetical designs, Design 6 is an extension of Design 5, experimentally supported, and performs better, suggesting that it would be beneficial for a cell to have all Clb/Cdk1 activities inhibiting SIC1 transcription. This finding is a testable prediction. Among the hypothetical designs that are not yet supported by experimental evidence, *Design 7*, which describes the inhibition of *CLB2* and *CLB3* syntheses by Sic1, exhibits the highest number of parameter sets in which limit cycles with 24 sampled limit cycles across 19 distinct phenotypes). This finding suggests the possible relevance of Sic1 transcriptional inhibition to guarantee a self-sustaining cell cycle, and is currently being tested in our laboratory.

As we exemplified for *Design 3*, we quantified the occurrence of parameters in the dominant terms (processes) for designs *4-9*, identifying phenotypes that (i) are valid, (ii) have a potential for oscillations, and (iii) yield limit cycles (Table 2.1 and Supplementary Code Repository). As we showed for *Design 3*, among the phenotypes that yield limit cycles, in designs *5-9* α_{yy} (Clb3 PFL) is the parame-

ter observed most frequently among the activatory regulations, followed by α_{yz} (Clb2/Cdk1 activation by Clb3/Cdk1) (Table 2.2). Intriguingly, when adding the inhibition of Clb5/Cdk1 by Clb2/Cdk1 in *Design 4*, the Clb2 PFL (α_{zz}) becomes the most dominant design. This likely reflects the crucial role of Clb2/Cdk1 in the modulation of *CLB5* synthesis, thus reinforcing the importance of positive feedback loops in the occurrence of sustained oscillations.

Furthermore, in all designs with the exception of *Design 3* and *Design 8*, both steps in the linear Clb cascade [32], α_{xy} (Clb5 $\rightarrow$ Clb3) and α_{yz} (Clb3 $\rightarrow$ Clb2), are more frequent than α_{xz} (Clb5 $\rightarrow$ Clb2), and in designs *5-9*, α_{yz} is equally or more frequent than α_{zz}. In fact, in all designs with the exception of *Design 4*, the Clb3 $\rightarrow$ Clb2 activation is the second dominant activatory regulation. Furthermore, once again, γ_{zx} (the APC-mediated inhibition of Clb5 by Clb2/Cdk1) and γ_{zy} (the APC-mediated inhibition of Clb3 by Clb2/Cdk1) are the parameters observed most frequently among the inhibitory regulations. The degradation rate of Sic1 from the Clb/Cdk1/Sic1 ternary complex (δ) was present in dominant terms of the majority of limit cycles for designs *3–9*. The same holds to a lesser extent for the degradation rate of the Clbs from those complexes (ϵ). Finally, the two novel interactions in *Design* 4 and *Design* 7 (inhibition of Clb5 synthesis by Clb2 and inhibition of Clb2 and Clb3 synthesis by Sic1 respectively) especially stand out in their contribution to the high number of limit cycles observed for these designs. For these two model designs the novel inhibitory interactions were dominant in nearly all identified limit cycles.

We additionally calculated the Pearson correlation coefficients between all parameter combinations across Designs *3–9* (Supplementary Information, Section 2.4 and Table S2.4). In line with the observations above, the parameters related to the Clb3 PFL and the APC-mediated inhibition of Clb3 by Clb2/Cdk1 are highly positively correlated in six out of the seven model designs (see Table S2.4). This is in line with the observation that both interactions occur as often dominant activatory and inhibitory terms, respectively (see Table 2.2). Intriguingly, even in *Design 4*, for which we observed a shift from the Clb3 PFL to the Clb2 PFL as the most often dominant activatory term in the limit cycles, this correlation remains high. This indicates that, even though the Clb3 PFL is more rarely dominant in this design, its strength still needs to be balanced by a Clb2/Cdk1-mediated inhibition.

Altogether, our findings highlight that the Clb3 and Clb2 PFLs, together with the linear cascade (Clb5 $\rightarrow$ Clb3 $\rightarrow$ Clb2) and the APC-mediated inhibitions driven by Clb2 are principles of design underlying a self-sustaining cell cycle network that may be conserved across evolution.

Limit cycles belong to distinct phenotypic regions spread across the parameter space and showcase a range of oscillation properties

Our analysis for designs *1–9* has retrieved hundreds of parameter sets that generate limit cycles (Table 2.1). A critical question to be addressed is whether these limit cycles belong to different, distinct regions in the parameter space. Part of

this question is addressed by the fact that, for each design, we have identified limit cycles across multiple phenotypes; this implies that different interactions and regulations are dominant across (some of) the limit cycles, and that they belong to distinct parameter space regions. However, design space phenotypes may overlap in the parameter space. A key concept here is that of the *robustness region*, i.e. the parameter space region around a limit cycle point within which parameters can be smoothly altered without interrupting the limit cycle behavior [49]. Generally, it is challenging to obtain a good approximation to a robustness region around a single point let alone compare multiple such regions for overlap.

To explore whether the identified limit cycles belong to non-overlapping robustness regions, we analyzed the spread of the parameter values through boxplots and by Principal Component Analysis (PCA) projection. For all designs, most parameter values cover multiple orders of magnitude (see Supplementary Information, Section 2.4 and Fig. S2.18). Interestingly, some parameters are consistently narrow in range across all designs: K_A, v_s and v_x (referring to the complex formation between Clb/Cdk1 complexes and Sic1, to the basal synthesis of Sic1 and to the basal synthesis of Clb5, respectively), indicating that they need to be tightly controlled in order to generate sustained oscillations. Conversely, α_{yy} (Clb3 PFL) is narrow in range in all designs except *Design 4*, and vice versa for α_{zz} (Clb2 PFL). This finding supports the observations of the flipped dominance of these parameters across the designs shown in Table 2.2. Similarly, in each design, either δ or ϵ show a narrow range. Interestingly, α_{yz} (Clb3 $\rightarrow$ Clb2), second step in the linear cascade (Clb5 $\rightarrow$ Clb3 $\rightarrow$ Clb2) shows a higher median value than the first step in all designs with the exception of *Design 4*, in agreement with its previously observed dominance.

Subsequently, we performed PCA (see "Methods" section 2.4) on the limit cycle parameter sets for designs *3–9* to visualize how the parameter values are spread throughout the 22 dimensional parameter space (Fig. 2.4). The limit cycles are spread across the two main principle components, suggesting that they are spread in the parameter space as well. For *Design 4*, many points appear to clump together; however, the scale on the axes is larger for this design than for the others, and many more limit cycles are found for this design which increase the overlap. This result is strengthened by the fact that for all designs the first two principle components never explain more than 61.9% of the variance in the data, indicating that there is also significant variance in the data along other orthogonal directions in the parameter space. For all designs in Fig. 2.4, limit cycles are separated both within and between phenotypes, indicating that, even within a single phenotype, limit cycles are found that are spread across different areas of the parameter space. The results illustrate the complex distribution of the parameter sets in the parameter space. A similar result was obtained by aggregating the limit cycles across designs *4–9* (bottom-right panel in Fig. 2.4). The aggregated results for designs *3–9* highlight that there are distinct areas of the parameter space that produce oscillations for different network designs.

Our analysis does not prove that the limit cycles do not belong to overlapping robustness regions (i.e. could potentially be found by continuation techniques), but indicates that the combined robustness region would have to cover

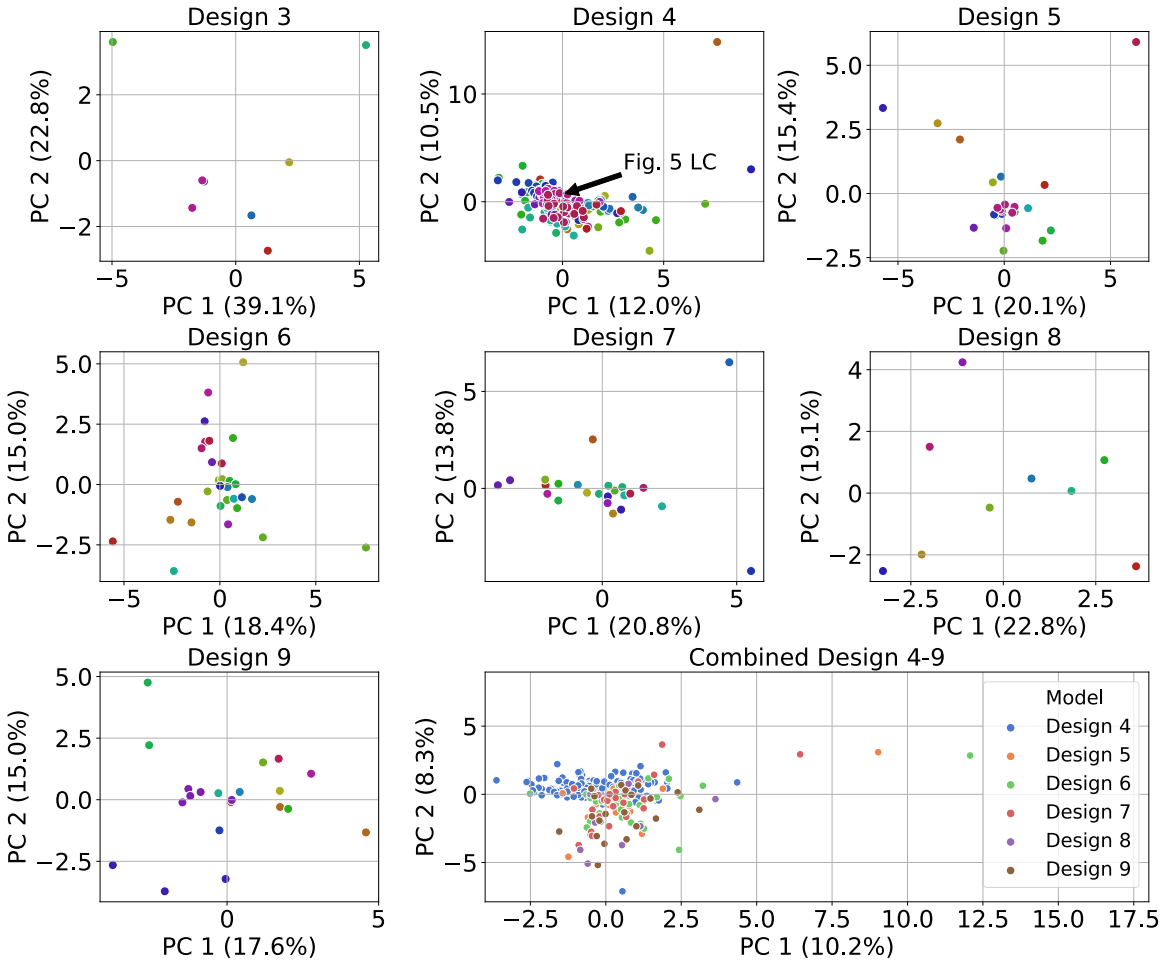


Figure 2.4: Projection of limit cycle parameter sets onto the first two principal component axes, for designs *3–9* separated and for designs *4–9* combined (bottom-right). Each dot represents a parameter set yielding a limit cycle. Parameter values were normalized to have a mean of zero and a standard deviation of 1 prior to principal component calculations, in order to deal with parameters spanning different orders of magnitudes. For the single panels, the colors indicate the unique phenotype each parameter sets belongs to. For the combined panel, the colors indicate the model design. For the purposes of this analysis, the unique inhibitory parameters in designs *4–9* were treated as the same parameter. The percentage of variance in the data explained by each principal component is listed on the axes. Values along the axes should be compared to the $[0, 1]$ unit interval since the principal components have a length of 1 and are linear combinations of the normalized parameters.

multiple orders of magnitude in most dimensions. Given that the vast majority of our parameter samples did not lead to limit cycles, such a vast robustness region seems unlikely. Altogether, Fig. 2.4 and S2.18 suggest that either some of our limit cycles belong to different robustness regions, or the robustness region must take an incredibly complex, unlikely and large N-dimensional shape.

We further quantified the differences in oscillation properties between the

limit cycles by looking at the period of the oscillation and at the minimal percentage of the oscillation amplitude with respect to the maximal concentration across species (see Supplementary Information, Section 2.4). In Fig. S2.19 and S2.20 boxplots of the period and amplitude (in terms of the minimum/maximum ratio), respectively, for designs *3–9* are shown. It can be observed that the limit cycles display a wide range in both properties; the wide range of parameter values covered between the limit cycles translates to differential oscillation properties.

As an illustrative example of the results that we obtained, we show a robustness analysis of a limit cycle of phenotype number 1906532 for *Design 4* (Fig. 2.5). We highlighted this limit cycle in the PCA plot for *Design 4* shown in Fig. 2.4; this limit cycle sits relatively close to most limit cycles found for *Design 4*. In Fig. 2.5A and 2.5B, a 2D slice of the parameter space across the K_{zx} - γ_{yy} plane (two dominant inhibitory regulations in this phenotype) is shown, where the phenotype number 1906532 (purple) borders with several other phenotypes (Fig. 2.5A) and the stability in terms of the number of non-negative eigenvalues may be observed (Fig. 2.5B). The black dot in Fig. 2.5A) and 2.5B) indicates a limit cycle that we identified – belonging to phenotype 1906532 –, which sits in an area where oscillations can be expected based on the eigenvalues. In Fig. 2.5C, properties of the limit cycle dynamics can be observed: (i) a period of about 27 minutes, resulting in a frequency of 0.037 minutes^{-1}; (ii) a similar (within 10-fold) order of magnitude of the amplitudes of the concentration of the four species considered (Sic1, Clb5, Clb3 and Clb2), consistent with previous observations [50, 51]; (iii) the correct temporal order of peaks of the four species considered: Sic1 in G1 phase, Clb5 in S phase, Clb3 in G2 phase, and Clb2 in M phase [10] and (iv) the amplitude of the oscillations covering most of the concentration range of the four species, i.e. their concentration sharply decreases and, in the case of total Clb3 and Clb2 becomes equal to zero when starting a new, successive cell cycle, as shown experimentally.

In Fig. 2.5D, a 1D bifurcation diagram is shown for the parameter K_{zx}, which refers to the transcriptional inhibition of Clb5/Cdk1 (x) by Clb2/Cdk1 (z) that is unique to *Design 4*. We observed that there is a range of K_{zx} values (in blue color) where sustained oscillations occur in the full model. The amplitude changes with the bifurcation parameter (indicated by the vertical size of the blue area). We observe a general agreement between the region in Fig. 2.5B, where the phenotypes support two positive eigenvalues, and the "robustness region" [49] of the limit cycle in Fig. 2.5D; however, it can be noted that the robustness region of the limit cycle is smaller than predicted by the eigenvalues of the phenotypic subsystem (yellow line). In Fig. 2.5E, a 2D parameter scan for the Clb3 PFL (α_{yy}) and the Clb2 PFL (α_{zz}) is shown in the form of what we consider as a "robustness region" . Fig. 2.5F shows a similar robustness region for the Clb2 PFL (α_{zz}) and the Clb3 NFL (γ_{yy}). These regions visualize how far the chosen parameters may be changed around the limit cycle such that the qualitative behavior does not change, and it is represented as a heatmap, where the color scale indicates the amplitude of oscillations. In this way, it is possible to investigate how mutations or environmental perturbations that lead to changes in model parameters can change or break the oscillatory behavior. This analysis highlights the complexity of combinations of parameters for which sustained oscillations may occur, and

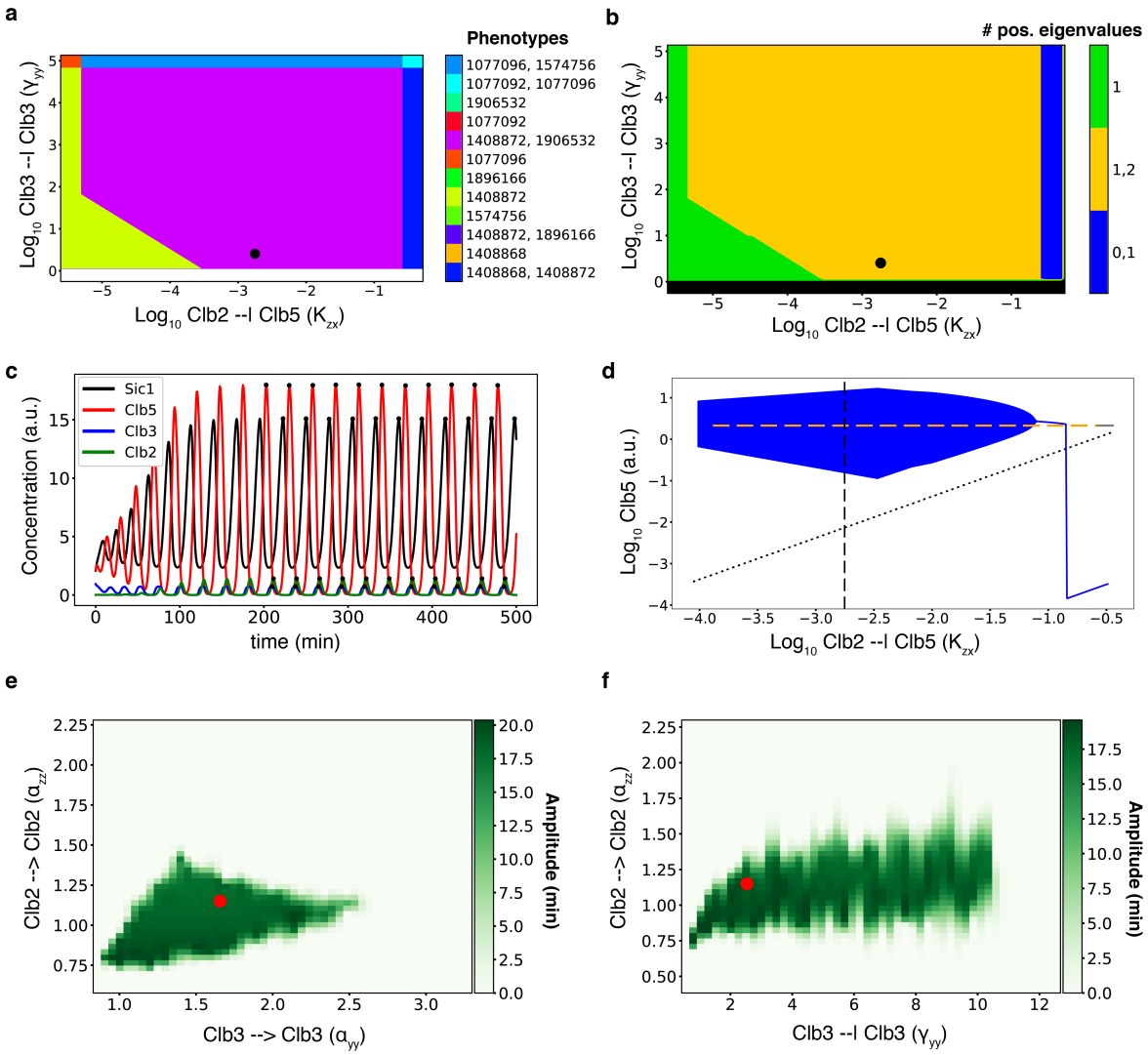


Figure 2.5: Robustness of a limit cycle for phenotype number 1906532 in *Design 4*. The black dot in (**a**) and (**b**), the time course in (**c**), the dotted line in (**d**) and the red dot in (**e**) and (**f**) represent the same limit cycle and parameter set. (**a**) 2D slice K_{zx} - γ_{yy} (with γ_{yy} representing the Clb3 PFL) of the design space visualized in regions corresponding to different phenotypes. White color represents absence of phenotypes, whereas the other colors indicate specific phenotypes, or combinations of overlapping phenotypes. Some regions relate to multiple phenotypes, indicating that these phenotypes overlap in this 2D projection of the parameter space. (**b**) The same 2D phase plane as in (**a**), visualizing the number of positive eigenvalues the steady-state of the phenotype. The orange area suggests potential for oscillatory behavior. Consistent with stability indications, limit cycles were not retrieved for most of the phenotypes in (**a**) based on our sampling (only phenotype 1574756 and 1906532) which fall in the orange area. The overlapping regions from (**a**) where multiple phenotypes are simultaneously valid in the same region of the design space are shown to support multi-stability. (**c**) Sustained oscillation time course of the limit cycle. (**d**) 1D bifurcation diagram plotting the oscillation amplitude (minimum and maximum of the oscillation) in

Figure 2.5: (Continued) the full model of the limit cycle for Clb5 (x on the y-axis and in the equations) while varying the K_{zx} parameter, is unique to *Design 4* and refers to the inhibition of Clb5/Cdk1 (x) by Clb2/Cdk1 (z). The blue line reflects the range of K_{zx} values yielding stable steady-state behavior, i.e. no oscillations, whereas the blue area represents the range of parameter values yielding oscillations, while keeping fixed all other parameter values. The yellow dashed line indicates an unstable steady-state with two non-negative eigenvalues in the phenotype. The grey dotted line indicates an unstable steady-state in the phenotype that has one non-negative eigenvalue, whereas the solid grey line indicates a stable steady-state; as there is more than one line, multi-stability occurs. (**e**) 2D robustness heatmap of the amplitude of oscillations (0 and white in case of a steady state) within a 2D slice of the parameter space. 1,000 random log-uniform samples of the parameters α_{yy} (Clb3 PFL) and α_{zz} (Clb2 PFL) were retrieved while keeping all other model parameters fixed. For each sample, the amplitude of oscillations in Clb5 is represented by the scale of green color. The radial basis interpolation algorithm (RBF) has been applied to infer the color for points in the plot that were not explicitly sampled. (**f**) Similar 2D robustness region heatmap as in (**e**) for α_{zz} (Clb2 PFL) and γ_{yy} (Clb3 NFL).

that, around the limit cycle identified by our pipeline (shown as a red dot), there is a region of the parameter space where oscillations are robust to change in these two parameters (albeit with a varying amplitude). This analysis highlights that this limit cycle is particularly sensitive to changes in the Clb2 PFL.

The analysis above can be used to identify phenotypes that are functional or dysfunctional based on specific features, e.g. in the tendency of a network design to oscillate. For instance, we can identify functional, healthy phenotypes that have a strong tendency to oscillate (2.5A), phenotype in violet color), as compared to dysfunctional phenotypes, which exhibit no oscillations (2.5A and 2.5B, all phenotypes except 1906532 and 1574756) or a reduced tendency to oscillate in only a small area of the parameter space, reflecting a reduced robustness of the limit cycle.

2.3 Discussion

Cell cycle networks are often modeled through checkpoint mechanisms, where the starting point of oscillations is reset upon reaching specific concentration thresholds of certain network components [20, 21]. To build computational cell cycle networks that oscillate autonomously is a challenge.

In this work we showed, to our knowledge for the first time, that some designs of the cell cycle network are suited to support truly autonomous oscillations, independent of checkpoint mechanisms, for a wide range of parameter sets. Specifically, we studied whether known or hypothetical designs may modulate the tendency to generate or stabilize sustained oscillations. We considered a minimal model of the network governing the activation of the mitotic

cyclin/Cdk1 (Clb/Cdk1) complexes in budding yeast [10], and analyzed 11 alternative network designs for their ability to yield limit cycles (Supplementary Information, Section 2.4 and Section 2.4). The model under investigation describes the sequence of events from the G1 through the M phases of the cell cycle and back to G1 again, assuming that to each phase is assigned one major functional component: Sic1, stoichiometric inhibitor of the Clb/Cdk1 complexes [52], to the G1 phase; Clb5/Cdk1, which promotes DNA replication dynamics [53], to the S phase; Clb3/Cdk1, which is involved in the spindle assembly [54], to the G2 phase; and Clb2/Cdk1, which promotes spindle formation and cell division [55, 56], to the M phase. In seven of these designs, from *Design 3* through *Design 9*, a quasi-steady-state approximation was introduced, which assumes an equilibrium between the Clb/Cdk1/Sic1 ternary complex and its free components, Clb/Cdk1 and Sic1. This assumption is novel in cell cycle models.

Our modeling effort, unlike existing cell cycle models that have been investigated in terms of their potential to show oscillations [8, 20, 21], supports the implicit hypothesis that there is a functional reason for autonomous oscillations. Aside from the novelty of our work in terms of the methodology that we have employed, we make a case for autonomous oscillations. The two opposing hypotheses of checkpoint models (i.e. the cell cycle system, by itself, should not favor oscillations, as these would disappear upon activation of checkpoints due to a cell's response to cellular damage, or to a not favorable response to environmental cues) and autonomous oscillations (i.e. the cell cycle system, by itself, should tend to exhibit self-sustained oscillations, independently from stimuli from the environment) are rather hard to prove or disprove. This is because oscillations do exist in living cells that may be due to cell cycle regulation alone, or to its interplay with the rest of the cell such as external factors, metabolic cues, etc. The fact that autonomously oscillating cell cycle models exist for mammalian cells [8], does not provide any strong support for either of the two hypotheses, because cell cycle models, overall, need to oscillate, and may be designed to do so. In the autonomous limit cycle models, limit cycle oscillations are identified by the presence of two complex conjugate eigenvalues with positive real part. In the checkpoint models, this property may or may not be present because – at specific points in the model dynamics – either the concentrations [20, 21] or both the concentration and the network wiring can be changed depending on the model under examination. The resetting of the position in the state space can force the model dynamics into a repetitive pattern that would not occur (in the same way) without the checkpoint(s).

The cell cycle models for budding yeast are currently incomplete: These models require the help of the modeler or a computer program to break and then restart the model at the end/beginning of every cycle. Our model is a limit cycle model, which cycles by itself without any periodic resetting. We developed a new methodology, based on initial work by Savageau and colleagues, to the point that we could scan the parameter space for possible limit cycles, by adding a search for complex positive eigenvalues around the Hopf bifurcations. This produced a type of model that then enabled us to determine which parameters control the occurrence and period of the cell cycling of yeast. Because the current cell cycle

models in budding yeast do not have a complete limit cycle, they cannot perform such a comprehensive control analysis.

Our cell cycle model structure was not designed to yield oscillations in general, but it can yield oscillations. Specifically, our work highlights that the underlying mechanisms of these oscillations are the Clb3-centered regulations – never considered in any available model of cell cycle regulation in budding yeast – which we have shown to exist in budding yeast cells [32]. The prediction that Clb3-centered regulations are the highest represented network motifs that lead to self-sustained, autonomous oscillations, provides a more subtle proof, and a much stronger evidence reconciling the checkpoint and autonomous oscillation views. Specifically, our results suggest that autonomous oscillations driven by Clb3/Cdk1 may occur when this complex is coupled and coordinated to the other S and M phase kinase complexes, Clb5/Cdk1 and Clb2/Cdk1, which are instead involved in the checkpoints. Whereas Clb5 and Clb2 have been described to be involved in the checkpoint mechanisms (Tyson and Novák's types of models), we propose that Clb3(/Cdk1) drives autonomous cell cycle oscillations to maintain cell proliferation. Clb3 being tightly coordinated together with Clb5 and Clb2, we envision that Clb3-mediated oscillations are maintained unless an activation of checkpoints terminates the autonomous oscillations.

Our findings do not rule out checkpoint mechanisms, but add an aspect that does not appear in pure checkpoint mechanisms, i.e. designs yielding autonomous oscillations. If we assume that only checkpoint mechanisms exist, then there would be no reason to expect such designs to occur in reality. However, the fact that they do occur supports the hypothesis that generating oscillations may provide an evolutionary advantage. Therefore, we hypothesize that "health" requires the capacity to oscillate autonomously – investigated in detail in this study – and the capacity to interrupt the oscillation due to checkpoint activation and/or unfavorable environmental cues – not investigated in this study and subject of a future modeling project.

It has been demonstrated [12] that including stochastic effects in checkpoint-based cell cycle models, using Langevin-type equations, can lead to qualitative changes in model dynamics and noise-induced oscillations. In this way, the dichotomy between checkpoint models and limit cycle oscillators is partially overcome by stochasticity, since checkpoints may be overcome through random fluctuations. Noise-induced oscillations also imply that inclusion of stochastic effects (in cell cycle models) may reshape and enlarge the regions in the parameter space that support oscillatory behavior. Experimental work regarding the role of noise in cell cycle regulation has been shown. Baumann and colleagues showed that stochastic telophase arrest of budding yeast mutants cannot be captured with a deterministic model, but can partially be captured by a stochastic model [57]. Similarly, by using a stochastic model of the G1/S transition, we showed that entrance into S phase is dependent on tight control of SIC1 mRNA transcription and degradation [58]. Moreover, Peccoud, Tyson and colleagues measured experimentally the size of fluctuations in mRNA levels of 16 proteins that are important in the cell cycle by using time-lapse fluorescence microscopy [59] and smRNA FISH [60, 61] and improved their cell cycle model to match the observa-

tions. Further development to incorporate the role of noise in cell cycle models in relation to sustained oscillations calls for software and methods supporting bifurcation analysis of stochastic differential equation models, also within the context of the System Design Space (SDS) methodology that we have used in this study or extensions thereof.

To study the dynamic effects of the designs over a wide range of parameter values, we applied the System Design Space methodology to analyze the phenotypes that the 11 network designs partitioned the parameters and state space into. These phenotypes can be associated to areas of the parameter space in which sustained oscillations in the form of limit cycles can occur. The ability to enumerate the phenotypic repertoire of each of the designs, and to explore the behavior of each phenotype in a model, allows for desired properties to be readily identified [27]. This, in turn, helps to reduce the computational effort by focusing the search of limit cycles on specific regions in the parameter space. The SDS methodology is therefore useful when studying natural systems and when engineering synthetic networks intended to be endowed with particular characteristics [62].

After applying our pipeline to identify oscillatory phenotypes (Fig. 2.2), we retrieved limit cycles for the network designs *3-9* but not for the designs *1A-2*. The latter lack the quasi-steady-state approximation, which is instead implemented in the former. The lack of observed oscillations in Design 1A provides an interesting case that is in line with the prevalent view that NFLs are required for sustained oscillations [15, 16]. Remarkably, the Clb3 positive feedback loop (Clb3 PFL) is recurrent in all network designs that yielded sustained oscillations with the exception of *Design 4*, where the Clb2 PFL takes over. Strikingly, PFLs have been shown to promote oscillations and switch-like responses that allow unidirectionality of cell cycle progression, by enhancing amplitude and robustness of cyclin/Cdk oscillations [5, 13, 14]. Our finding that PFLs are important for obtaining sustained oscillations in the cell cycle is in agreement with these previous studies. In recent work, Novák and colleagues highlighted the importance of the PFL between SBF and Cln1,2 for the cell size checkpoint in G1 phase by using checkpoint models [63]. Similarly, they highlighted the importance of two antagonistic PFLs of Cdk1:CycB and PP2A:B55 for interphase-M phase transitions in the mammalian cell cycle, by using a non-checkpoint model that exhibits bistability and hysteresis [64]. Contrarily to these two studies, our work focused on autonomous oscillations rather than on multiple different steady-state attractors. A novel insight from our work is that it appears that our models do not require both the Clb3 and Clb2 PFLs but, depending on the absence or presence of the inhibition of Clb5/Cdk1 by Clb2/Cdk1 (designs *3, 5–9* vs. *Design 4*), one PFL has a more stabilizing effect on the oscillations than the other. The general result that we retrieve concerning the NFLs and PFLs is in agreement with the literature. However, the specific (combinations of) PFLs and NFLs that are found in oscillating phenotypes and parameter sets (see Table 2.2) were not predictable a-priori.

Furthermore, the regulatory activation Clb3 $\rightarrow$ Clb2 (α_{yz}), which forms the linear *CLB* cascade [32] together with Clb5 $\rightarrow$ Clb3 (α_{xy}), is observed more frequently than Clb5 $\rightarrow$ Clb2 (α_{xz}). Therefore, our analyses point to the relevance of

Clb3-centered regulations for the generation of sustained Clb/Cdk1 oscillations in budding yeast. Importantly, our results for *Design 4* and *Design 5* support the experimental evidence that the inhibitory regulations, Clb2/Cdk1 on Clb5 in *Design 4* and Clb2/Cdk1 on the Clb/Cdk1 inhibitor Sic1 in *Design 5*, play a crucial role in cell cycle regulation.

Among the hypothetical network designs, *Design 7* is of particular interest because it rationalizes a recent experimental observation for which the molecular mechanisms remain at the moment obscure. This design describes the inhibition of *CLB2* and *CLB3* syntheses by Sic1, as an attempt to describe the experimental evidence that Sic1 oscillations rescue viability of cells with low levels of mitotic Clb cyclins [48]. Our findings indicate that including this inhibitory regulation resulted in a higher number of limit cycles as compared to other hypothetical network designs. This result suggests that Sic1 inhibition on the synthesis of *CLB2* and *CLB3* may be relevant to guarantee a self-sustained cell cycle. We speculate that this inhibitory regulation might occur through a physical interaction of Sic1 on transcription factors that drive synthesis of these mitotic cyclins, such as Fkh2, which we have described to be the regulator driving both *CLB2* and *CLB3* transcription [32]. The mammalian counterpart of Sic1, the cyclin/Cdk inhibitor $p27^{Kip1}$ [65], has indeed been shown to behave as a transcriptional repressor, by binding to and inhibiting a number of gene promoters through E2F4/p130 complexes [66]. The specific role of $p27^{Kip1}$ as transcriptional repressor is to recruit G1 cyclin/Cdk complexes needed for p130 phosphorylation in early-mid G1 phase [67]. This regulation of a cyclin/Cdk inhibitor at gene promoters is unknown in budding yeast, and a direct involvement of Sic1 as transcriptional repressor, potentially through Fkh2, calls for a detailed experimental investigation, which we are currently conducting in our laboratory.

Importantly, we addressed that the limit cycles we identified belong to multiple different phenotypic regions, in the System Design Space sense (which sometimes partially overlap), and that the parameter values cover multiple orders of magnitude and are spread across a PCA projection, whose variance is spread across many principal components. Our analysis suggests that it is improbable that all limit cycles found for each particular model design belong to the same robustness region. However, it is currently not possible to prove that their robustness regions do not overlap. In order to answer this question, it would be useful to explore whether methods that approximate single robustness regions [49] can be applied to multiple limit cycles to prove their separation in the parameter space.

Our work shows how the SDS methodology can aid in the identification of qualitatively distinct behavior of complex systems, resulting in phenotypes that are characterized by a tendency to generate oscillations within a definite network design. In this respect, phenotypes that exhibit oscillations can be considered functional phenotypes of a cell, which exhibit oscillations or a high number of oscillations as compared to dysfunctional phenotypes, which may exhibit no or low amount of oscillations. This approach may represent an interesting avenue of further research to embed the missing details of the minimal cell cycle network considered in this work into existing checkpoint models. The cell cycle network

incorporates designs that are particularly suited to support sustained oscillations, suggesting that this biochemical process may have evolved to generate or stabilize autonomous oscillations, at least under some conditions.

We envision that populations of cells consist of subpopulations expressing different phenotypes, and that individual cells are able to dynamically shift their network configurations so as to effectively alter their phenotype. This would allow evolutionarily selectable differences between subpopulations to emerge. Point-mutations and shifts in gene expression, e.g. up- and down-regulation of inhibitors, provide valid mechanisms by which the functioning of any network interaction could be altered. Such alterations can impact the strength of, or entirely block, a network interaction, e.g. a PFL. For example, in our cell cycle networks, binding or phosphorylation affinity of the Clb/Cdk1 complexes could be altered. Consequently, cells may theoretically be able to dynamically shift network configurations, as we have proposed for metabolic networks [68, 69] (also see Ch. 5), causing switches in phenotype. Our results indicate that, if these changes occur within the core cell cycle regulatory network, an impact on the ability of the network to exhibit oscillations can be observed. Therefore, differences in the affinity of Clb/Cdk1 complexes to bind and phosphorylate Fkh transcription factors and, vice versa, in the affinity of Fkh to the CLB promoters may be expected. These scenarios are currently investigated in our laboratory.

Given the evolutionary conservation of the cell cycle network across eukaryotes, our approach may be translated to human cell cycle models, in which components are often mutated in disease [70]. In the mammalian cell cycle there does not exist a one-to-one relationship between the robustness and maintained frequency of the cell cycle of individual cells on the one hand, and the "health" of the whole organism on the other hand. For human cells, the whole organism may not be "healthy" when each cell would cycle robustly, or with a higher frequency. Here our approach reverses: in the context of cancer, our analysis could highlight network design principles that would be good (healthy) for the particular disease state (the cancer). Regardless, applied to the whole organism or to populations of cells within the organism that may become deregulated, as in cancer, it is of interest to identify network properties that result in changes in the robustness and frequency of oscillations. Thus, our pipeline can be used to point to precise molecular strategies of intervention to restore molecular designs that may be disrupted in disease.

2.4 Methods

All Python and MATLAB code, Jupyter Notebooks and COPASI files are available as part of a Github repository (`https://github.com/barberislab/Autonomous_minimal_cell_cycle_oscillator`).

Simulation of Ordinary Differential Equation (ODE) models

Time course analyses were conducted in MATLAB 2017a by using the ode15s solver or in Python 2.7 by using the scientific Python (SciPy) version 1.1.0, Numeric Python (NumPy) v1.14.1, Design Space Toolbox Python module v0.3.0a4 and the related C toolbox v0.3.0a6. In the Python scripts, a sequence of integrators was set up in case that one of the methods would fail to integrate accurately. The sequence implemented was the following: lsoda, bdf and dopri. The Python and MATLAB code used to generate all our analyses are provided in the Supplementary Code Repository.

Application of the SDS Toolbox

The Design Space Toolbox V2 for Python 2.7 [31] was used to apply the System Design Space methodology to the 11 network designs considered in this work. The functionality of the toolbox was first tested by reproducing previously published results by Savageau and colleagues [27]. Subsequently, their pipeline to analyze the phenotypic repertoire for oscillatory phenotypes was implemented in Python. The novelty of our implementation is two-fold: (i) the generalization of this pipeline, to be applicable to any predefined kinetic model in terms of a set of equations in SciPy notation; (ii) the approach of Savageau and colleagues was improved to include extensive parameter sampling within the boundaries of phenotypes characterized by potential oscillations (oscillatory phenotypes), i.e. areas in the parameter space with two non-negative eigenvalues. The pipeline of our work is shown in Fig. 2.2. The pipeline consists of a Python script containing model definitions, a set of newly written Python functions that are wrapped around the Design Space Toolbox V2, and several Jupyter notebooks [71] that analyze a model and properties of any given limit cycle, respectively. All files are available in the Supplementary Code Repository. The main Jupyter notebook reads the model to be analyzed (defined in GMA form), and then proceeds to (i) set up the design space for the model, (ii) identify all valid phenotypes, (iii) identify the stability of each valid phenotype indicated by the presence of two complex conjugate eigenvalues with non-negative real part, by sampling each valid phenotype for a user defined number of times (in this study 250 parameter samples were collected), and (iv) retrieve limit cycle behavior by integrating the full kinetic model in the GMA form (Supplementary Information, Section 2.4) for the sampled parameter set. The other Jupyter notebooks allow users (i) to analyze properties of a limit cycle, specifically to draw: 1D bifurcation diagrams, phenotype phase planes, stability diagrams and robustness regions for a user defined set of limit cycles and bifurcation parameters; (ii) to reproduce the results from Table 2.1 and Table 2.2; and (iii) to reproduce the boxplots of the period, amplitude, parameter values and the PCA plots.

Parameter space sampling to find oscillatory phenotypes

The phenotypes characterized by a network design may theoretically exhibit oscillatory as well as steady-state behaviors for a range of parameter values. To de-

termine the steady-state(s) of a phenotype, a representative point within the phenotypic region of the parameter space must be found. The Design Space Toolbox can determine valid parameter sets for a given phenotype. The log-linear boundaries associated with the boundary conditions for the system (referred to as an S-system) representing a particular phenotype define a continuous subspace of parameter values. Within this space, the terms in the S-system dominate the neglected terms and describe the dominant behavior (Supplementary Information, Section 2.4. These boundaries enable linear programming problems to be solved to identify a set of parameter values at a vertex of the phenotypic region.

We implemented a sampling approach in two steps. Since steady-state stability may change within a phenotype when model parameters are altered, sampling just once may not give an accurate view of a phenotype's stability. We sampled parameter sets for each valid phenotype of a particular network design 250 times. For each sample of the parameter values, we first used the functionality of the System Design Space Toolbox to determine the steady-state and the presence of non-negative real part complex conjugate eigenvalues of the steady-state for the combination of a given phenotype and parameter set. Second, for valid phenotypes satisfying the necessary condition for sustained oscillations, i.e. two complex conjugate non-negative eigenvalues at the fixed point, we determined the dynamics in the full model in the GMA form, using the previously calculated steady-state as the initial condition, and checked for the presence of a limit cycle. Different initial conditions may give rise to different attractors and, hence, iterating this procedure for multiple initial conditions may result in the identification of more limit cycle attractors. However, in this work, we did not take this approach.

To sample a valid parameter set for a given phenotype, we first used the *valid_parameter_set* function in the Design Space Toolbox to obtain a valid parameter set for the phenotype. We then rearranged the parameters in the model by shuffling them into a random so as to avoid biased sampling due to the fact that the sample for each parameter may alter the phenotypic boundaries for the following parameters. Subsequently, for each parameter in the randomized order, we (i) determined the phenotypic tolerance: 1D boundaries of the phenotype when keeping all other parameters the same (utilizing the *vertices_1D_slice* function in the System Design Space Toolbox), and (ii) log-uniformly sampled the range of numbers between these boundaries. This sequence of steps ensures that we retrieve a set of unique parameter sets that are specific for a given phenotype, and random. We opted for log-uniform random sampling due to inherent problems that we observed with uniform random sampling. In uniform random sampling in 1D, ranges with the same length have the same probability of being sampled. When a phenotype has, for example, a range of $[0, 100]$, 99% of the samples will fall in the $[1, 100]$ interval and 10% will fall in the [90, 100] interval. This has the consequence that parameter sets with relatively low parameter values are exceedingly rare, especially when sampling multiple parameters simultaneously as is commonly the case with biochemical models. This problem is further aggravated by the fact that the effects of parameters in the model are multiplicative, rather than additive. As do Metabolic Control Analysis (MCA) and Biochemical

Systems Theory (BST), we think that equal relative changes are equally important; hence, we concluded that log-uniform sampling was appropriate for this work. In all model designs, parameters were limited to the range $[10^{-9}, 1000]$.

To check whether sampled parameter sets yield limit cycles, and not just damped oscillations, we integrated the system of ODEs for the full model in the GMA form (Supplementary Information, Section 2.4) in a series of subsequent time windows. After each successive time window we first checked whether the integration of the ODEs proceeded successfully without error. Second, we used the last time window to check the properties of the time course for each of the model species (Sic1, Clb5, Clb3, Clb2) by identifying all maxima that: (i) were within 5% of the global maxima in the current time window. When five such ordered maxima for each species occurred in the time course, we considered the time course to exhibit sustained oscillations. A limit cycle trajectory exhibits multiple, successive peaks in a repeating pattern and would therefore satisfy the aforementioned criteria. The five ordered maxima of each species may represent a yeast cell dividing at least five times. To accept the time course as a limit cycle, we additionally required that: (i) all species have an oscillation amplitude of at least 10% of their global maximum, (ii) the ratio between the global maxima across all species is less than 100-fold (experimentally, there is less than a three-fold difference in the concentration of the four species considered in the model [50, 51], and (iii) the identified maxima in step (ii) are not only found in the beginning of the time course, since this would otherwise indicate a damped oscillation. The main results presented in Table 2.2 did not change when we required an oscillation amplitude of at least 50% (data not shown). After each successive time integration window, we checked that the conditions above were met. If the limit cycle conditions were satisfied, the integration was stopped. Finally, as last criteria, we required the identified limit cycle oscillation to exhibit the correct cell cycle order (Sic1, Clb5, Clb3, Clb2). Conversely, if the conditions were not satisfied, the integration was continued unless a steady-state had been reached. We defined a steady-state as when none of the concentrations changed more than 1% of their global maxima in the last time window. If the time course did not exhibit a limit cycle or a steady-state within 10,000 minutes, the integration was stopped and we concluded that the time course did not exhibit oscillations. In our hands, these two tests are sufficient to identify limit cycles. In rare cases, this approach may erroneously detect a limit cycle although there is in fact a slowly decaying, damped oscillation or slowly increasing oscillations; however, such cases should become clearer from inspection of the time course.

Principal component analysis

PCA projects a set of N-dimensional vectors along a new coordinate axis specified by orthogonal and uncorrelated vectors which are ranked according to how much of the variance in the data they explain. These "principal components" are linear combinations of the original parameters and are of unit length. Each principal component is associated with a percentage of the variance in the original dataset that it explains. As a result we can visualize features (parameter values in this

case) in two-dimensional space in such a way that parameter sets that are “close together" (i.e. not showing much variation) will appear together on the PCA plot.

Observations that are far apart in the PCA plot are separated in the original space along the directions specified by the principal components if there was significant deviation within the original dataset to begin with. However, parameter sets that are far apart in the original data may sometimes cluster closely together on the PCA plots if the first principal components do not explain a lot of the variance in the dataset.

We first normalized the parameter sets by dividing by the standard deviation and subtracting the mean so that each parameter has a mean of 0 and a variance of 1 across the different limit cycle parameter sets before applying the PCA algorithm. This helps avoid bias from the multiple orders of magnitude covered in the parameter values.

Supplementary Information

Model derivations and initial limit cycles for *Design 1A* through *Design 3*

Here, we introduce model designs *1A-3* which are updated and expanded versions of the minimal cell cycle model of Barberis and colleagues [10], shown in Fig. S2.1. We will start by discussing the original model.

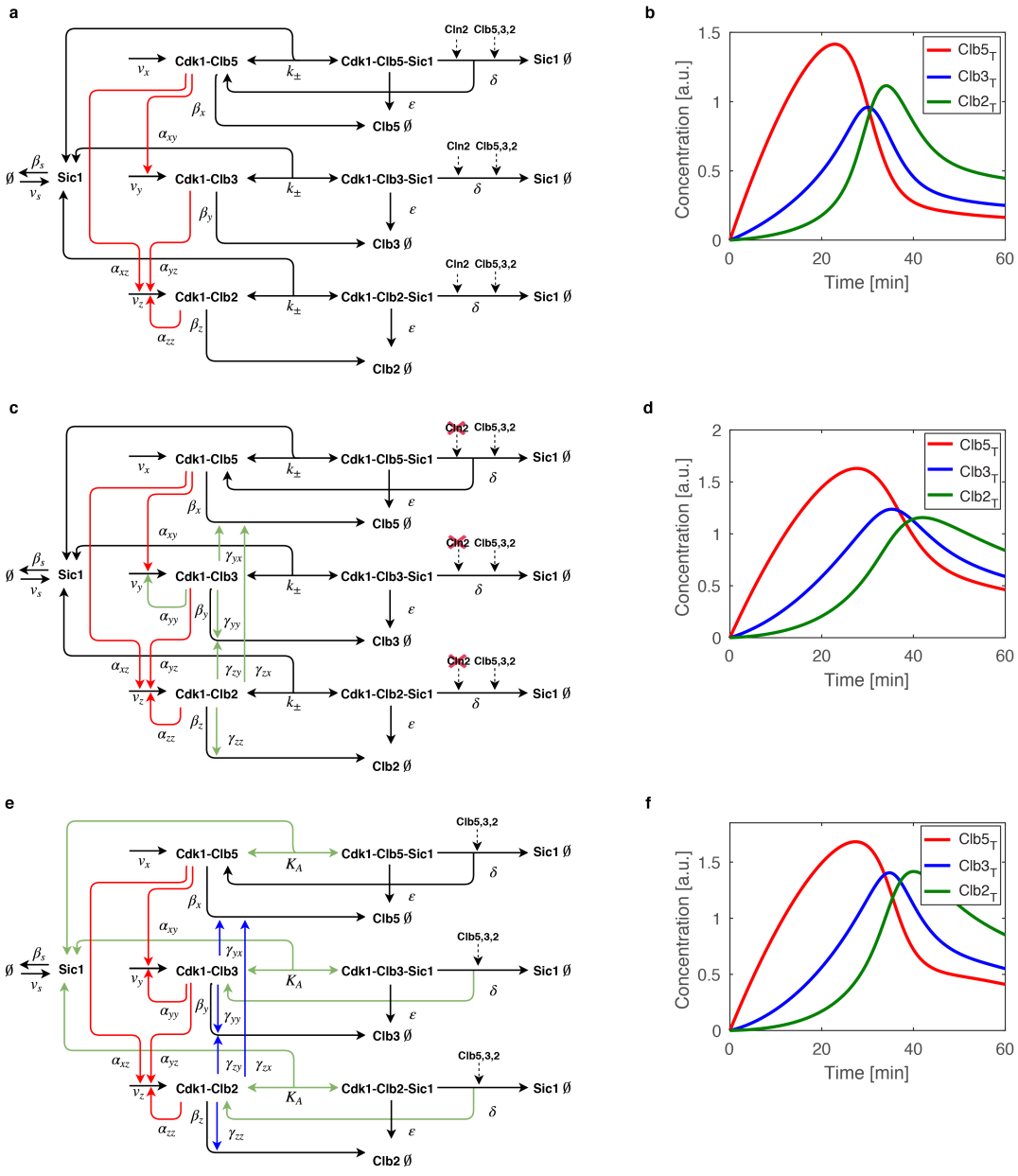


Figure S2.1: Alternative designs of a minimal cell cycle network in budding yeast. Model schemes and simulated time courses of waves of Clb/Cdk1 complexes are

Figure S2.1: (Continued) represented as follows: *Design 1A* (**a**) and its time course (**b**); *Design 1B* and *Design 1C* (**c**) and simulated time course for *Design 1C* (**d**); *Design 2* and *Design 3* (**e**) and simulated time course for *Design 3* (**f**). First, the minimal cell cycle model (*Design 1A*) was expanded to generate *Design 1B* (see Section 2.4) by including: (i) a positive feedback loop (PFL) by Clb3/Cdk1 on *CLB3* (Clb3 PFL; α_{yy}) in **c**) through a transcription factor [32], (ii) a negative feedback loop (NFL) of Clb3/Cdk1 on itself through activation of degradation (Clb3 NFL; γ_{yy} in **c**), the four known Clb-regulated degradations (inhibitory regulations) mediated by both the Clb/Cdk1 and (ii) complexes and Anaphase Promoting Complex (APC) (γ_{zz}, γ_{zy}, γ_{zx} and γ_{yx}) (see Fig. 1C in [45] for details). *Design 1C* removes the Cln2/Cdk1-mediated contribution to Sic1 degradation from the Clb/Cdk1/Sic1 ternary complexes (red crosses in **c**). *Design 2* includes the salvaging of Clb3/Cdk1 and Clb2/Cdk1 upon degradation of Sic1 from the Clb/Cdk1/Sic1 ternary complexes (see green lines in **e** indicated with δ) (see Section 2.4 such salvaging was already in place for Clb5/Cdk1 upon degradation of the Cdk1-Clb5-Sic1 ternary complex in **a** and **c**). *Design 3* describes the complex formation between Sic1 and Clb cyclins by the association equilibrium constant (K_A), instead of the forward (k_+) and backward (k_-) parameters in *Design 2*, as a consequence of the quasi-steady-state approximation, i.e. assumes a fast equilibrium between the free forms of Sic1 and Clb/Cdk1 complexes and the ternary complexes they form, which is reflected by a high K_A value. This implies that steady-state concentration of Clb/Cdk1/Sic1 ternary complexes changes with Sic1 dynamics, and that the fraction of free Clb cyclins is directly related to free Sic1 at any moment (Supplementary Information, Section 2.4). This assumption is supported by *in vitro* experimental evidence indicating a strong binding of Sic1 to the Clb/Cdk1 complexes. (**a**, **c** and **e**) Red arrows indicate activation of each *CLB* gene, thereby of each Clb/Cdk1 complex, by any previous Clb/Cdk1 complex in the cascade; blue arrows indicate APC-mediated Clb inhibition by Clb/Cdk1 complexes, resulting in Clb degradation; black arrows indicate all other reactions; and green arrows indicate the progressive changes with respect to the previous kinetic model(s). Dotted arrows indicate the Cln(/Cdk1)- and Clb(/Cdk1)-mediated phosphorylation of Sic1 in Clb/Cdk1/Sic1 ternary complexes, resulting in its degradation. (**b**, **d** and **f**) Time courses representing the total concentrations of Clb5, Clb3 and Clb2 as function of time are shown in red, blue and green, respectively (for sake of clarity, Sic1 time course has been omitted). Time courses were obtained by simulating the kinetic models with the canonical parameter set (see Table S2.1).

Notation and assumptions in the minimal cell cycle model

The starting point is our published minimal cell cycle model [10]. We consider a system of seven species: (i) three representing the complexes that Cdk1 forms with the three pairs of B-type cyclins, Clb5,6, Clb3,4 and Clb1,2, which we refer to as x, y and z, and (ii) the inhibitor Sic1, which we refer to as s, that binds and inhibits all three Clb/Cdk1 complexes ($s{\cdot}x$, $s{\cdot}y$ and $s{\cdot}z$). In living cells of budding

yeast there is a distinction between the roughly constant concentration of Cdk1 and the varying concentration of Clb cyclins throughout cell cycle progression. This results in varying concentrations of the Clb/Cdk1 complexes; however, this distinction is not needed for the modeling purposes presented here.

The mathematical description of the minimal model is in terms of a system of coupled Ordinary Differential Equations (ODEs). Specifically, the model considers: (i) basal degradation of all four species in their free form; (ii) basal synthesis of the Clb/Cdk1 complexes; (iii) forward activation in the Clb cyclin cascade from one Clb/Cdk1 complex to another through phosphorylation of a transcription factor; (iv) backward inhibition in the Clb cascade from one Clb/Cdk1 complex to another through the Anaphase-Promoting Complex (APC); (v) reversible formation of the ternary complex formed by any of the Clb/Cdk1 complexes and Sic1; (vi) degradation of Sic1 in any of the Clb/Cdk1/Sic1 ternary complexes; and (vii) degradation of Clb cyclins in the Clb/Cdk1/Sic1 ternary complexes. We consider x, y and z to be active in the binary complexes, and inactive in the ternary complexes with s. We also consider that degradation of Sic1 can occur when Sic1 is either free or bound to Clb/Cdk1 in the Clb/Cdk1/Sic1 complexes. We refer to reference [10] and Table S2.2-S2.3 for the experimental evidence of the interactions described above.

We assume basal synthesis (v) of each Clb/Cdk1 complex to be at a constant rate, and basal protein degradation (β) to be proportional to the current concentration of a species. We further implement that activation (α) of one Clb/Cdk1 complex by the previous one occurs through activation of a transcription factor, and is thus assumed to be proportional to the activating Clb/Cdk1 complex. Inhibition (γ) from a Clb/Cdk1 complex to the previous one is considered to occur through the APC, and is thus assumed to be proportional to the product of both species involved. Complex formation (k^+) and dissociation (k^-) are proportional to the concentrations of the species forming the complex and to the concentration of the complex, respectively. Complex dissociation due to Sic1 or Clb degradation from the Clb/Cdk1/Sic1 ternary complexes (δ and ϵ, respectively) is assumed to be proportional to the complex concentration. In our mathematical notation, we will use brackets to denote concentrations and the $A \cdot B$ notation to denote a complex formed by two species A and B. In the equations, we neglect to mention Cdk1, as it is most abundant in living cells as compared to the Clb cyclins that activate it, thus being the limiting species.

The Barberis 2012 model

In terms of the notation introduced above, the model from [10] may be represented as follows:

$$
\begin{aligned}
\frac{d[x]}{dt} &= v_x - \beta_x[x] - \gamma_{yx}[x][y] - \gamma_{zx}[x][z] - k_x^+[s][x] + k_x^-[s \cdot x] \\
&\quad + \delta_x \left(1 + [x] + [y] + [z]\right)[s \cdot x] \\
\frac{d[y]}{dt} &= v_y - \beta_y[y] + \alpha_{xy}[x] - \gamma_{zy}[z][y] - k_y^+[s][y] + k_y^-[s \cdot y] \\
\frac{d[z]}{dt} &= v_z - \beta_z[z] + \alpha_{zz}[z] - \gamma_{zz}[z]^2 + \alpha_{xz}[x] + \alpha_{yz}[y] - k_z^+[s][z] + k_z^-[s \cdot z] \\
\frac{d[s]}{dt} &= -\beta_s[s] - (k_x^+[x] + k_y^+[y] + k_z^+[z])[s] + k_x^-[s \cdot x] + k_y^-[s \cdot y] + k_z^-[s \cdot z] \\
\frac{d[s \cdot x]}{dt} &= k_x^+[s][x] - k_x^-[s \cdot x] - \delta_x \left(1 + [x] + [y] + [z]\right)[s \cdot x] - \epsilon_x[s \cdot x] \\
\frac{d[s \cdot y]}{dt} &= k_y^+[s][y] - k_y^-[s \cdot y] - \delta_y \left(1 + [x] + [y] + [z]\right)[s \cdot y] - \epsilon_y[s \cdot y] \\
\frac{d[s \cdot z]}{dt} &= k_z^+[s][z] - k_z^-[s \cdot z] - \delta_z \left(1 + [x] + [y] + [z]\right)[s \cdot z] - \epsilon_z[s \cdot z]. \qquad (2.1)
\end{aligned}
$$

The canonical parameter set from the Barberis 2012 model

Table S1 shows the parameter notation from the original model and the translation to the notation used in this text. The parameter values are those used in the original publication [10].

Parameter	Value	Parameter	Value	Parameter	Value
v_x (k_1)	0.1	α_{yz} (k_B)	1	ϵ_x (k_4)	0.01
v_y (k_7)	0.01	α_{zz} (k_D)	0.1	ϵ_y (k_17)	0.01
v_z (k_9)	0.001	γ_{yx} (k_E)	0.7	ϵ_z (k_13)	0.01
β_x (k_6)	0.7	γ_{zx} (k_F)	0.7	k_x^+ (k_2)	5
β_y (k_8)	0.7	γ_{zy} (k_G)	0.7	k_y^+ (k_{15})	5
β_z (k_10)	0.7	γ_{zz} (k_H)	0.7	k_z^+ (k_{11})	5
β_s (k_{26})	0.001	δ_x (k_5)	0.05	k_x^- (k_3)	0.5
α_{xy} (k_A)	1	δ_y (k_{18})	0.05	k_y^- (k_{16})	0.5
α_{xz} (k_C)	0.1	δ_z (k_{14})	0.05	k_z^- (k_{12})	0.5

Table S2.1: Parameters of the Barberis 2012 model as originally published. The parameter names refer to those used in this work, with the matching parameter from the original publication in brackets. Note that some parameters might seem to have different values than in the original publication. This is due to the fact that, originally, the positive regulations were multiplied by basal synthesis, and the negative regulations were multiplied by basal degradation. Here, we incorporate these multiplications into the equations. All parameters have units of min^{-1} when we adopt the convention that concentrations are dimensionless.

Design 1

In this work, we start from a model that builds on the Barberis 2012 model (Fig. S2.1A) but contains more regulations in less parameters. We assume that several parameters are equal across molecular species, i.e. all δ, ϵ, k^+, and k^- parameters are assumed to be equal, allowing us to describe the same regulations with less parameters. This means that we assume rates of complex formation and dissociation, Sic1 degradation from the ternary complexes and total degradation of ternary complexes to be equal among all the three species. Of note, this procedure reduces the number of parameters from 27 as in the original model to 19, but the model output for the default parameter set in [10] remains the same, since these parameters were already assumed to be equal (see Table S2.1).

Recently, Fkh2 was identified as a transcription factor of *CLB3* transcription, and that Clb3 may activate Clb2 [32]. Given that Clb3 may phosphorylate Fkh2, a positive feedback loop through Fkh2 has been envisioned in the model. In the Barberis 2012 model, negative feedback inhibitions among the Clb/Cdk1 complexes were considered; among these, the self-inhibition of M phase cyclins. However, in the model this regulation was only implemented for Clb2. Here, we extend the model to additionally implement the Clb3 self-inhibition. Consequently, our model has two new parameters ($\alpha_{\mathrm{yy}}, \gamma_{\mathrm{yy}}$), increasing the total number of parameters to 21.

The Barberis 2012 model included a term representing the Cln1,2/Cdk1 phosphorylation on Sic1 for its degradation, when the latter is in the Clb/Cdk1/Sic1 ternary complexes. This term was incorporated as a constant parameter, because time-varying concentrations of Cln1,2/Cdk1 were not considered. However, this leads to an issue in the units of the corresponding parameter, which was simultaneously used to indicate the phosphorylation of Sic1 mediated by the Clb/Cdk1 complexes. For the former, δ should have units of 1/time and, for the latter, units of 1/(concentration*time). For this reason, and for simplicity, we now introduce this term as a separate parameter λ (increasing the total number of parameters to 22) with units of 1/time and with the convention that $\delta = \lambda$.

Of note, the Barberis 2012 model did not include any synthesis of Sic1, instead assuming that Sic1 level is high at the start of a cell cycle and decays throughout the cell cycle, rising again during the M phase to restart the cycle. This means that, by design, the model could not generate multiple oscillations in time since this requires resetting of the Sic1 concentration. However, in this work we are interested to retrieve sustained oscillations in all four molecular species, and for this the synthesis term for Sic1 is required. Therefore, we added the simplest possible synthesis term for Sic1 in the form the of the parameter v_s which has units of concentration/time. This increases the total number of parameters to 23. With the new parameters just introduced, the set of equations describing the model can be written as follows:

$$
\begin{aligned}
\frac{d[x]}{dt} &= v_x - \beta_x[x] - \gamma_{yx}[x][y] - \gamma_{zx}[x][z] - k^+[s][x] + k^-[s \cdot x] \\
&\quad + \delta\left([x] + [y] + [z]\right)[s \cdot x] + \lambda[s \cdot x] \\
\frac{d[y]}{dt} &= v_y - \beta_y[y] + \alpha_{xy}[x] + \alpha_{yy}[y] - \gamma_{zy}[z][y]\gamma_{yy}[y]^2 - k^+[s][y] + k^-[s \cdot y] \\
\frac{d[z]}{dt} &= v_z - \beta_z[z] + \alpha_{zz}[z] + \alpha_{xz}[x] + \alpha_{yz}[y] - \gamma_{zz}[z]^2 - k^+[s][z] + k^-[s \cdot z] \\
\frac{d[s]}{dt} &= v_s - \beta_s[s] - k^+\left([x] + [y] + [z]\right)[s] + k^-\left([s \cdot x] + [s \cdot y] + [s \cdot z]\right) \\
\frac{d[s \cdot x]}{dt} &= k^+[s][x] - k^-[s \cdot x] - \delta\left([x] + [y] + [z]\right)[s \cdot x] - \epsilon[s \cdot x] - \lambda[s \cdot x] \\
\frac{d[s \cdot y]}{dt} &= k^+[s][y] - k^-[s \cdot y] - \delta\left([x] + [y] + [z]\right)[s \cdot y] - \epsilon[s \cdot y] - \lambda[s \cdot y] \\
\frac{d[s \cdot z]}{dt} &= k^+[s][z] - k^-[s \cdot z] - \delta\left([x] + [y] + [z]\right)[s \cdot z] - \epsilon[s \cdot z] - \lambda[s \cdot z]. \qquad (2.2)
\end{aligned}
$$

Of note, the Barberis 2012 model does not salvage Clb3,4/Cdk1 and Clb1,2/Cdk1 when Sic1 is degraded in the Clb/Cdk1/Clb ternary complexes, whereas it does recover Clb5,6/Cdk1, as experimentally demonstrated [72]. This can be seen in the equations from the δ term occurring only for x and not for y and z. Last, upon Clb degradation from the Clb/Cdk1/Clb ternary complexes, Sic1 is not recycled, as it can be seen from the equation for the evolution of $[s]$, since no terms with ϵ occur. Biologically, this means that we assume that Sic1 is not available to function, i.e. it is inactivated due to re-localization to the nucleus.

Since we have the ability to 'turn off' certain interactions by setting the corresponding parameters to 0, here we investigate three different scenarios for *Design 1*: *A*, *B* and *C*. For each scenario, we show a transient oscillation ($v_s = 0$), a limit cycle ($v_s \neq 0$) and we analyze the control on the period of the oscillations.

Finding initial parameter sets and analyzing period control coefficients

COPASI [73] was used to find initial parameter sets that yielded sustained cyclin/Cdk oscillations. The slider functionality of COPASI allows for the manual adjustment of one or several model parameters, and for the visualization of the output of a given combination of parameter sets. Alternatively, limit cycles may be found by using the *Manipulate* function in Mathematica or the bifurcation software such as MATCONT [74, 75] and XPPAUT [76]. The COPASI files for designs 1A-3 are available as Supplementary Code Repository.

Once a set of parameters that generated oscillatory behavior had been identified, the control exerted by the model's parameters on the period of the oscillation was be analyzed. Control can be quantified by control coefficients: logarithmic derivatives of system properties, e.g. fluxes and concentrations, with respect to kinetic parameters. Summation laws have been derived about control coefficients for parameters with dimension 1 \time with respect to both autonomous and forced oscillations [77, 78]. Taking log-log derivatives of a period (or any other

derivative of a stationary state function) is analogous to the concept of a control coefficient for a steady state flux or concentration i.e. $C_p^\tau = \frac{\partial \log(\tau)}{\partial \log(p)}$ [79, 80]. The sensitivity analysis on limit cycles by control coefficients was conducted by using the PeTTSy toolbox in MATLAB [81]. The model equations in MATLAB were converted in the specific format required by PeTTSy; furthermore, files containing parameters and initial conditions were defined as specified in the PeTTSy manual. With these three files (equations, parameters, initial conditions) the model can be read by PeTTSy. The PeTTSy input files are available as part of the Supplementary Code Repository. Once a model has been imported into PeTTSy, a new parameter set may be defined that generates oscillations. PeTTSy then integrates the equations and returns a time course. At this point, the user may run the *Derivatives* function and will be prompted to accept or reject the solution. According to the PeTTSy documentation it is crucial, for accuracy reasons, to accept only solutions with a $\log[\text{condition number}] < 36$. We used 400 time blocks to guarantee accuracy for designs *1A-3*. After calculating the derivatives, the sensitivity of the period with respect to parameter changes was analyzed.

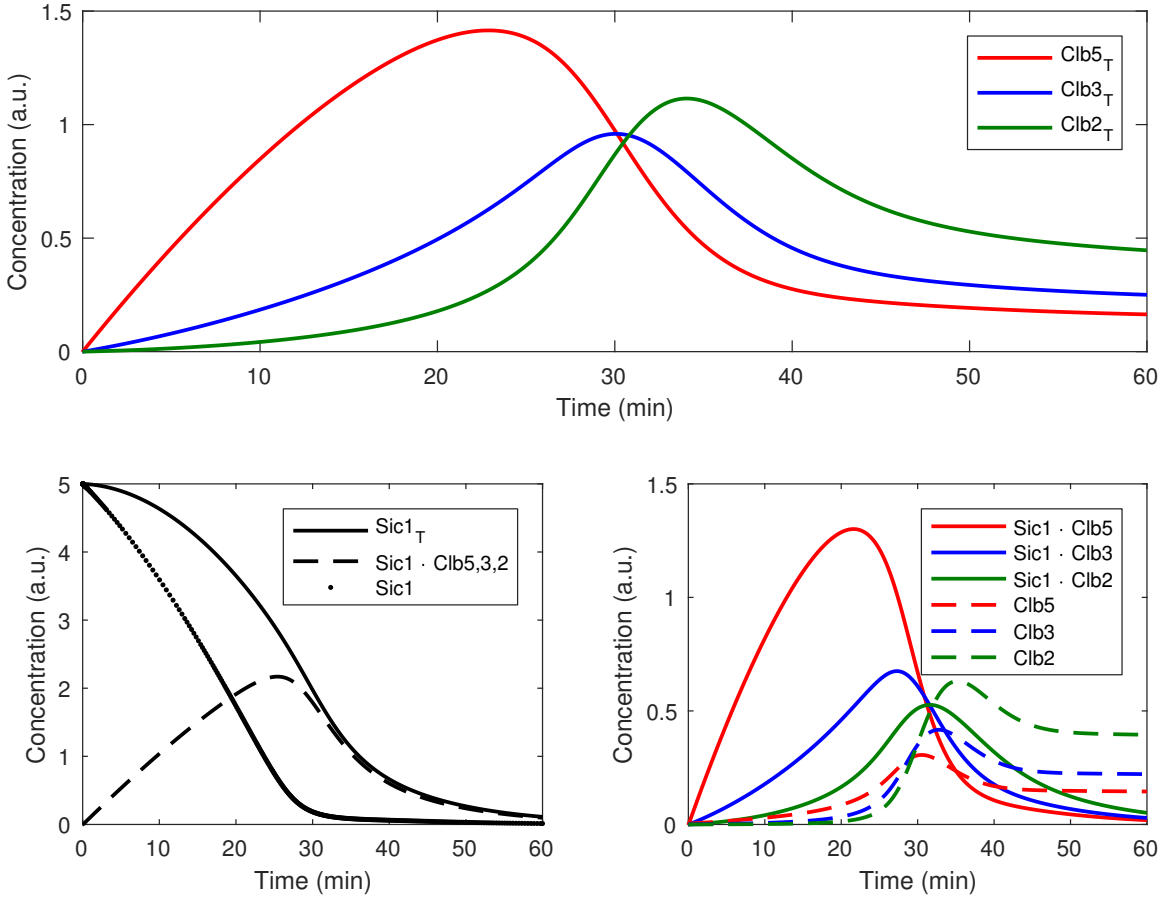


Figure S2.2: Time courses for *Design 1A*. (Top) time courses for the Clb5/Cdk1, Clb3/Cdk1 and Clb2/Cdk1 total concentrations. (Bottom-left) time courses for total Sic1, ternary Sic1 complex with Clb/Cdk1, and free Sic1. (Bottom-right) time courses for the binary (Clb/Cdk1) and ternary (Clb/Cdk1/Sic1) complexes.

Design 1A: No inhibition through the APC

To start with, we considered the model presented in Fig. 2 of [10] (Fig. S2.1A). In Fig. S2.2, the evolution over one cell cycle is plotted by using the equations introduced above, and for the parameter values specified in Table S2.1, with all γ parameters, α_{yy} and v_s set equal to 0. The top panel of Fig. S2.2 shows a transient oscillation in the total Clb/Cdk1 concentrations that exhibits their sequential rise and (near simultaneous) fall over time. Of note, waves of the total concentrations arise predominantly due to the concentrations of the Clb/Cdk1/Sic1 ternary complexes, although waves can be also observed for the concentration of Clb/Cdk1 complexes (Fig. S2.2 bottom-right). With regard to the parameter set, cyclin degradation was considered to be of several orders of magnitude faster when in complex with Cdk1 alone (β) as compared to degradation from the Clb/Cdk1/Sic1 ternary complexes (ϵ). Biologically, this translates to a small likelihood of cyclin degradation from the ternary complexes.

This model is able to exhibit limit cycles when a shift in the parameter set occurs. When we simply turn on the synthesis of Sic1 by setting v_s equal to 0.3 dampened transient oscillations occur. Sustained oscillations are found when increasing the rate of ternary complex formation from 5 to 20 (see Fig. S2.3); the sustained oscillation is a limit cycle with a period of roughly 113 minutes.

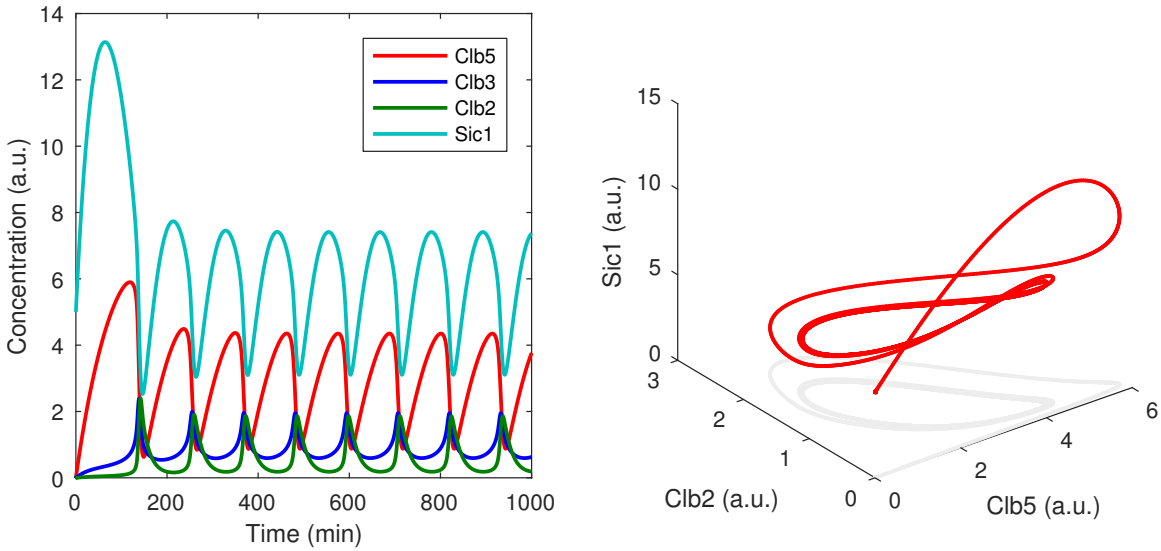


Figure S2.3: Sustained oscillations in *Design 1A*. (Left) Time courses of the four total species concentrations. (Right) 3D view of the limit cycle in the Clb5-Clb2-Sic1 space.

Sensitivity analysis of the period of oscillations points to a few key parameters controlling the period, namely: β_x, ϵ, v_s, λ, v_x. The former three parameters yield positive derivatives, whereas the latter two parameters yield negative derivatives (see Fig. S2.4). Intriguingly, these parameters controlling the period are all either basal synthesis or basal degradation rates for various species in the model. Moreover, the parameters referring to the regulations among the Clb/Cdk1 complexes do not control significantly the period length.

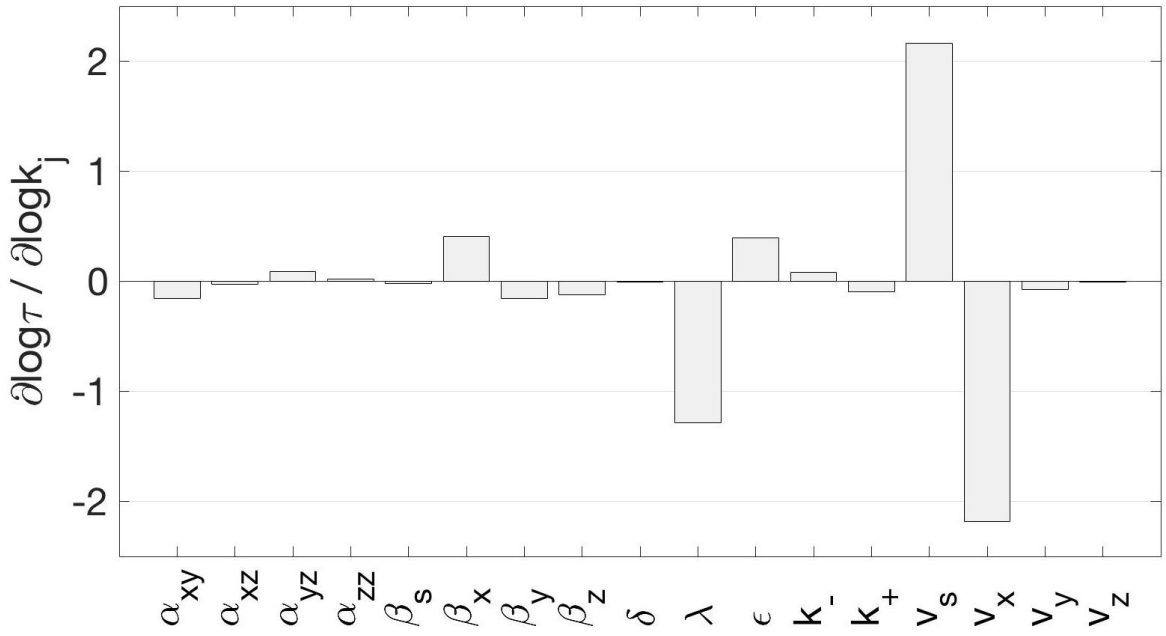


Figure S2.4: Logarithmic period derivatives for *Design 1A*.

***Design 1B*: Regulatory inhibitions and Clb3 positive and negative feedback loops**

The *Design 1B* includes the Clb/Cdk1-mediated inhibition on Clb/Cdk1 complexes and the potential of Clb3 self-activation, as we have recently proposed [32] (Fig. S2.1B). We also include the potential of Clb3 self-inhibition, similarly to that identified on Clb2/Cdk1 [82]. We assume, as we consistently do, that self-activation and self-inhibition are proportional to the protein concentration, in this case y. For this extended model, we set the inhibitory parameters (γ) to a non-zero value, and add self-activation and self-inhibition terms to the Clb3 (y) ODE, where we assume the new parameters to have the values $\alpha_{yy} = 0.1$ and $\gamma_{AB} = 0.7$, i.e. all Clb/Cdk1-mediated inhibitions have the same parameter value as in the Barberis 2012 model (Table S2.1). In Fig. S2.5, the time course for the canonical parameter is plotted for *Design 1B*. Of note, the peaks of total Clb/Cdk1 concentrations appear slightly earlier as compared to *Design 1A*.

Activating Sic1 synthesis and increasing the Clb/Cdk1/Sic1 ternary complex formation rate, i.e. using the same parameter set as for *Design 1A*, we again find sustained oscillations, now with a period of roughly 95 minutes (see Fig. S2.6). Of note, this mirrors the anticipation of the Clb/Cdk1 waves in the transient oscillations by a decrease in the limit cycle period.

We repeated the sensitivity analysis on the period of the sustained oscillations (see Fig. S2.7), which show the same qualitatively results as compared to *Design 1A*. Interestingly, none of the newly added inhibitory parameters (γ's) nor α_{yy} have much control over the period of oscillations.

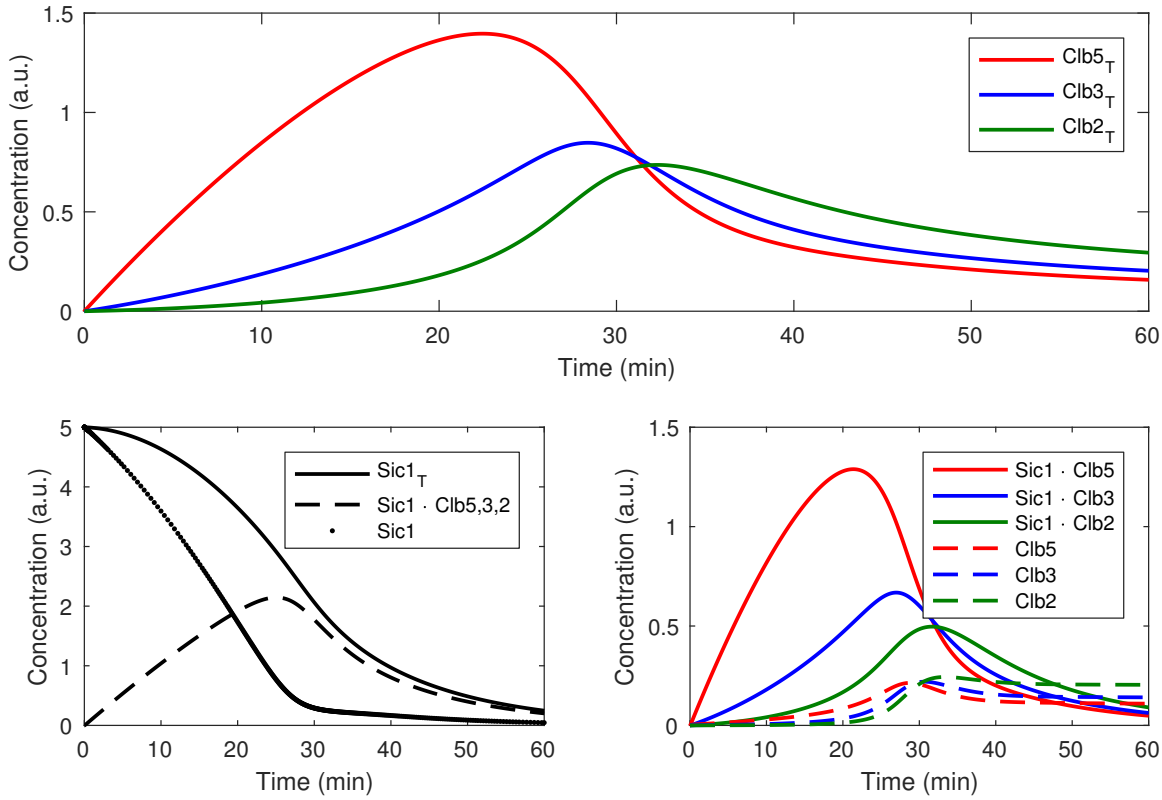


Figure S2.5: Time courses for *Design 1B*. (Top) time courses for the total concentrations of the three Clb/Cdk1 complexes: Clb5/Cdk1, Clb3/Cdk1 and Clb2/Cdk1. (Bottom-left) time courses for total Sic1, Sic1 in complex with Clb/Cdk1 complexes, and Sic1. (Bottom-right) time courses for the binary (Clb/Cdk1) and ternary (Clb/Cdk1/Sic1) complexes.

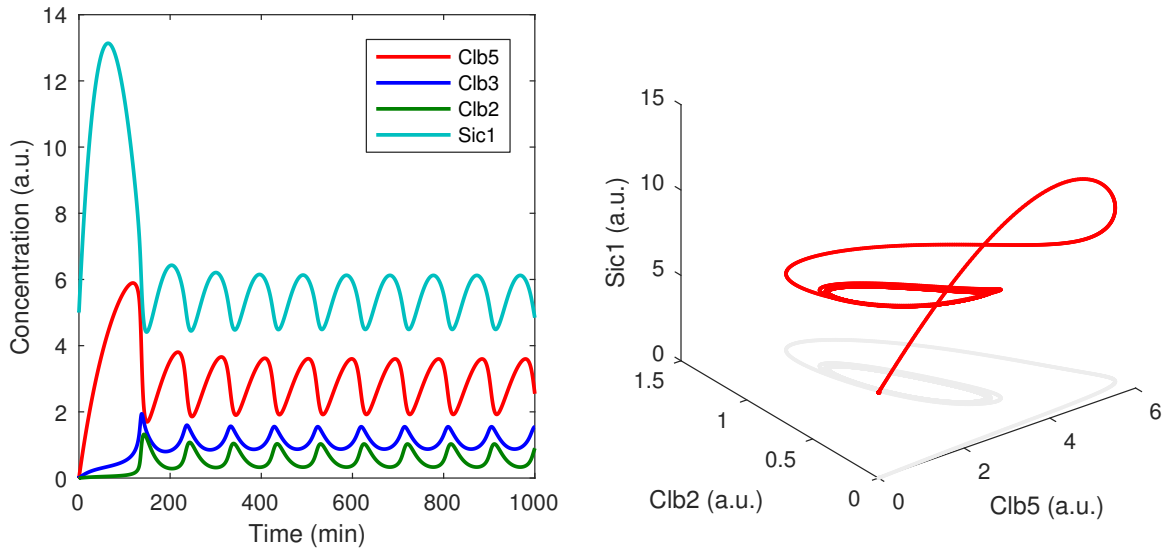


Figure S2.6: Sustained oscillations in *Design 1B*. (Left) Limit cycle for the total concentrations of the four species. (Right) 3D view of the limit cycle in the Clb5-Clb2-Sic1 space.

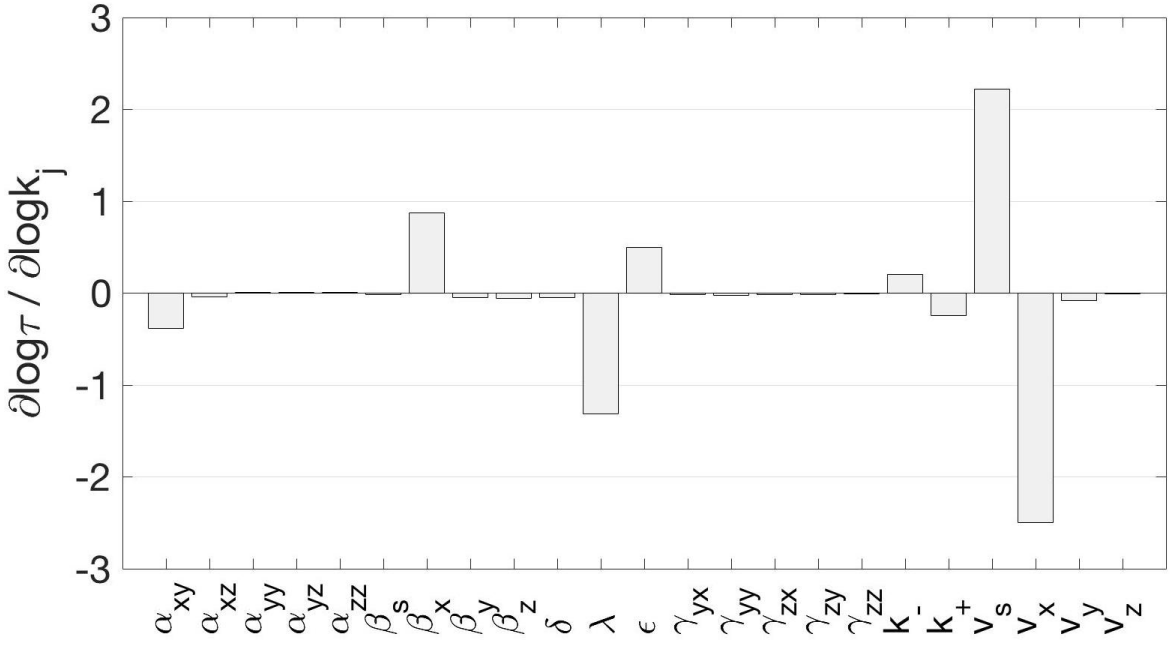


Figure S2.7: Logarithmic period derivatives for *Design 1B*

***Design 1C*: Neglecting Cln1,2/Cdk1 on Sic1 degradation from Clb/Cdk1/Sic1 ternary complexes**

To simplify the model further, we neglect the λ parameter, i.e. $\lambda = 0$, which was added in the Barberis 2012 model as a way to include basal levels of Sic1 degradation from the Clb/Cdk1/Sic1 ternary complexes (possibly due to Cln1,2/Cdk1) (red crosses in Fig. S2.1B). However, this parameter does not change the structure of the model, and its effect can be compensated for by ϵ and δ. In Fig. S2.8, the transient oscillations for the canonical parameter is plotted for *Design 1C*. Of note, the peaks of total Clb/Cdk1 concentrations appear later as compared to *Design 1B*.

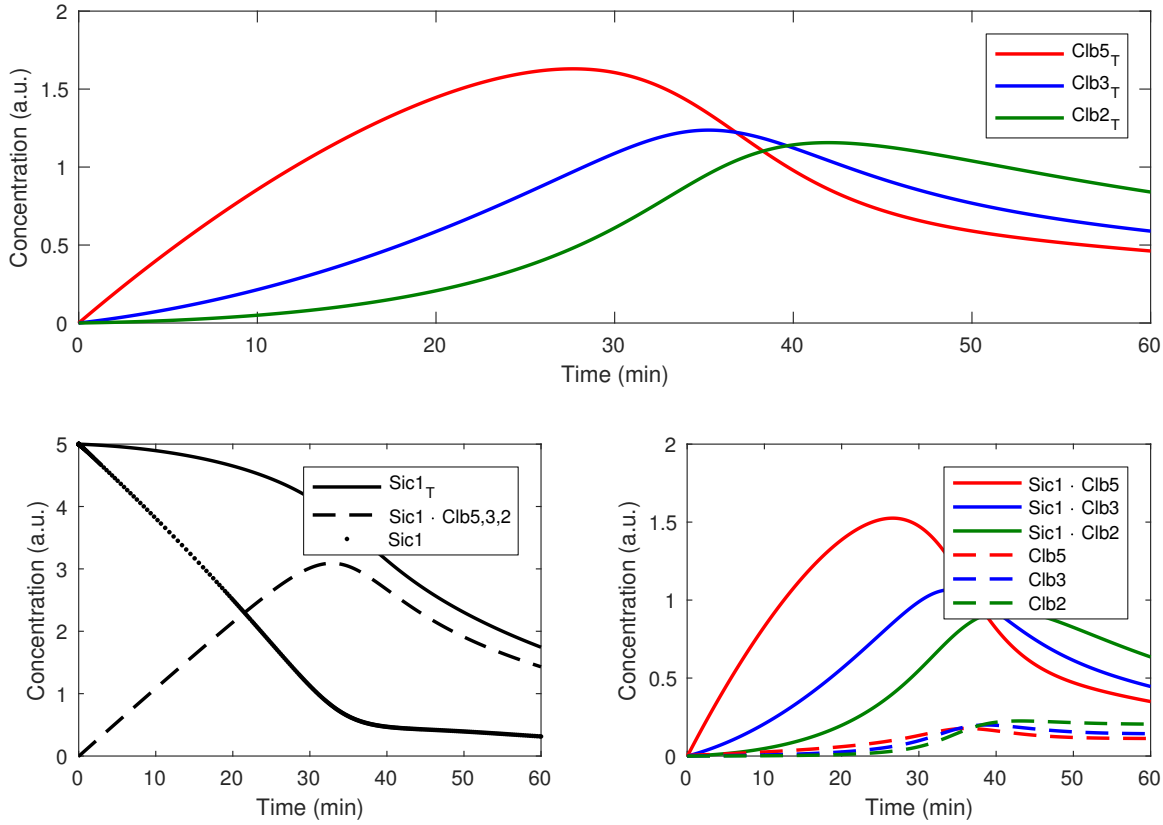


Figure S2.8: Time courses for *Design 1C*. (Top) time courses for the total concentrations of the three Clb/Cdk1 complexes: Clb5/Cdk1, Clb3/Cdk1 and Clb2/Cdk1. (Bottom-left) time courses for total Sic1, Sic1 in complex with Clb/Cdk1 complexes, and Sic1. (Bottom-right) time courses for the binary (Clb/Cdk1) and ternary (Clb/Cdk1/Sic1) complexes.

To obtain sustained oscillations in *Design 1C*, a number of parameters had to be altered. Ultimately, autonomous oscillations were obtained varying the following parameters as compared to the limit cycle parameter set for *Design 1A* and *Design 1B*: $v_s = 0.18$, $\beta_s = 0.003$, $\delta = 0.1$, $\lambda = 0$, $\epsilon = 0.005$. The resulting limit cycle is plotted in Fig. S2.9, and has a period of roughly 96 minutes.

The sensitivity analysis of the sustained oscillations is shown in Fig. S2.10. Some changes may be observed as compared to *Design 1B*. There is a significant

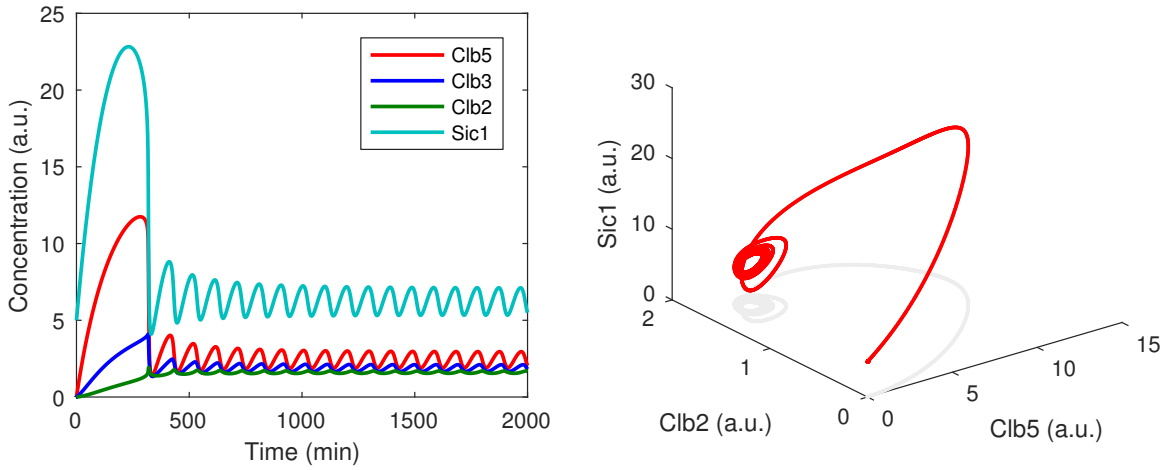


Figure S2.9: Sustained oscillations in *Design 1C*. (Left) Limit cycle for the total concentrations of the four species. (Right) 3D view of the limit cycle in the Clb5-Clb2-Sic1 space.

increase in the absolute value of the control of most parameters (note the difference in y-axis values as compared to Fig. S2.7) but especially the control coefficients for v_x, v_s, β_x, δ, α_{xy} and α_{yz} increase in absolute value. This may be understood by the fact that λ had significant control in *Design 1A* and *Design 1B* and, in its absence, other parameters have to 'take over' this control. It appears that the δ parameter 'takes over' the control that was previously of λ, the other parameters that are changed in terms of their control. Therefore, there may be some compensation mechanism for the absence of basal Sic1 degradation from the Clb/Cdk1/Sic1 ternary complex that is not dependent on the free Clb/Cdk1 concentrations.

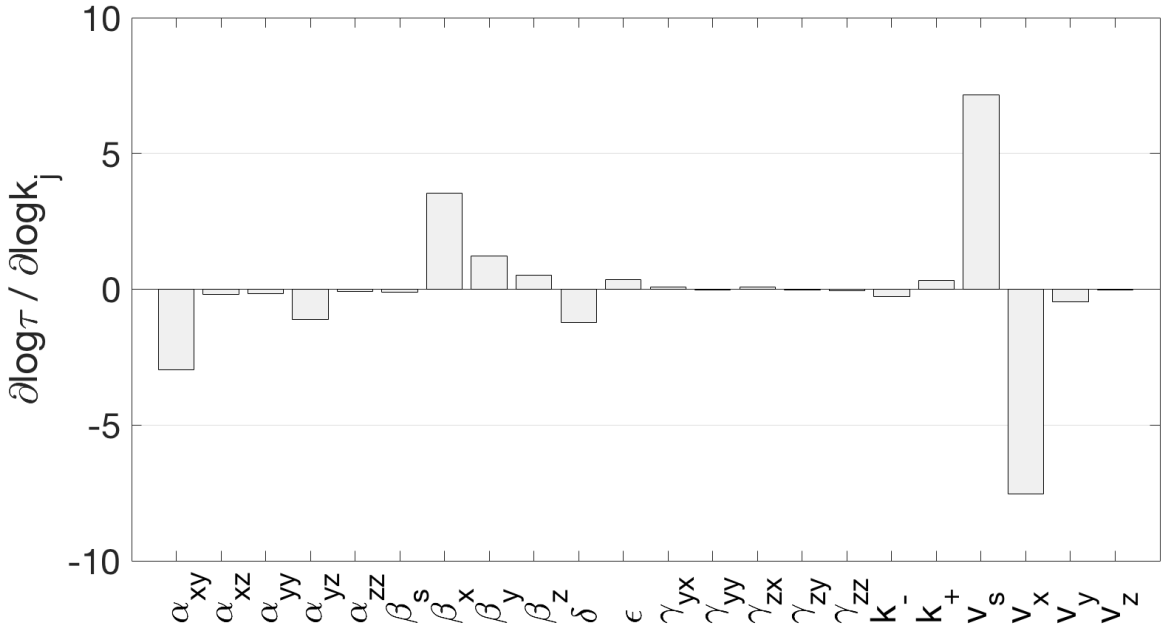


Figure S2.10: Logarithmic period derivatives for *Design 1C*. Of note, the control in λ has switched to parameter δ, and the control for v_x, v_s, β_x and α_{xy} has significantly increased as compared to *Design 1B*.

Design 2: Including Clb3/Cdk1 and Clb3/Cdk1 salvaging

Design 2 includes the salvaging of Clb3/Cdk1 and Clb2/Cdk1 complexes when Sic1 is degraded from the Clb/Cdk1/Sic1 ternary complexes, as previously demonstrated for Clb5/Cdk1 [72] (Fig. S2.1C) The updated equations are as follows:

$$
\begin{aligned}
\frac{d[x]}{dt} &= v_x - \beta_x[x] - \gamma_{yx}[x][y] - \gamma_{zx}[x][z] - k^+[s][x] + k^-[s\cdot x] \\
&\quad + \delta\left([x]+[y]+[z]\right)[s\cdot x] \\
\frac{d[y]}{dt} &= v_y - \beta_y[y] + \alpha_{xy}[x] + \alpha_{yy}[y] - \gamma_{zy}[z][y] - \gamma_{yy}[y]^2 - k^+[s][y] + k^-[s\cdot y] \\
&\quad + \delta\left([x]+[y]+[z]\right)[s\cdot y] \\
\frac{d[z]}{dt} &= v_z - \beta_z[z] + \alpha_{zz}[z] + \alpha_{xz}[x] + \alpha_{yz}[y] - \gamma_{zz}[z]^2 - k^+[s][z] + k^-[s\cdot z] \\
&\quad + \delta\left([x]+[y]+[z]\right)[s\cdot z] \\
\frac{d[s]}{dt} &= v_s - \beta_s[s] - k^+\left([x]+[y]+[z]\right)[s] + k^-\left([s\cdot x]+[s\cdot y]+[s\cdot z]\right) \\
\frac{d[s\cdot x]}{dt} &= k^+[s][x] - k^-[s\cdot x] - \delta\left([x]+[y]+[z]\right)[s\cdot x] - \epsilon[s\cdot x] \\
\frac{d[s\cdot y]}{dt} &= k^+[s][y] - k^-[s\cdot y] - \delta\left([x]+[y]+[z]\right)[s\cdot y] - \epsilon[s\cdot y] \\
\frac{d[s\cdot z]}{dt} &= k^+[s][z] - k^-[s\cdot z] - \delta\left([x]+[y]+[z]\right)[s\cdot z] - \epsilon[s\cdot z].
\end{aligned}
\tag{2.3}
$$

Of note, positive δ terms in the ODEs for y and z have been added. The transient oscillations for the canonical parameter set for *Design 2* are shown in Fig. S2.11 The differences with the transient oscillations observed for *Design 1C* are negligible.

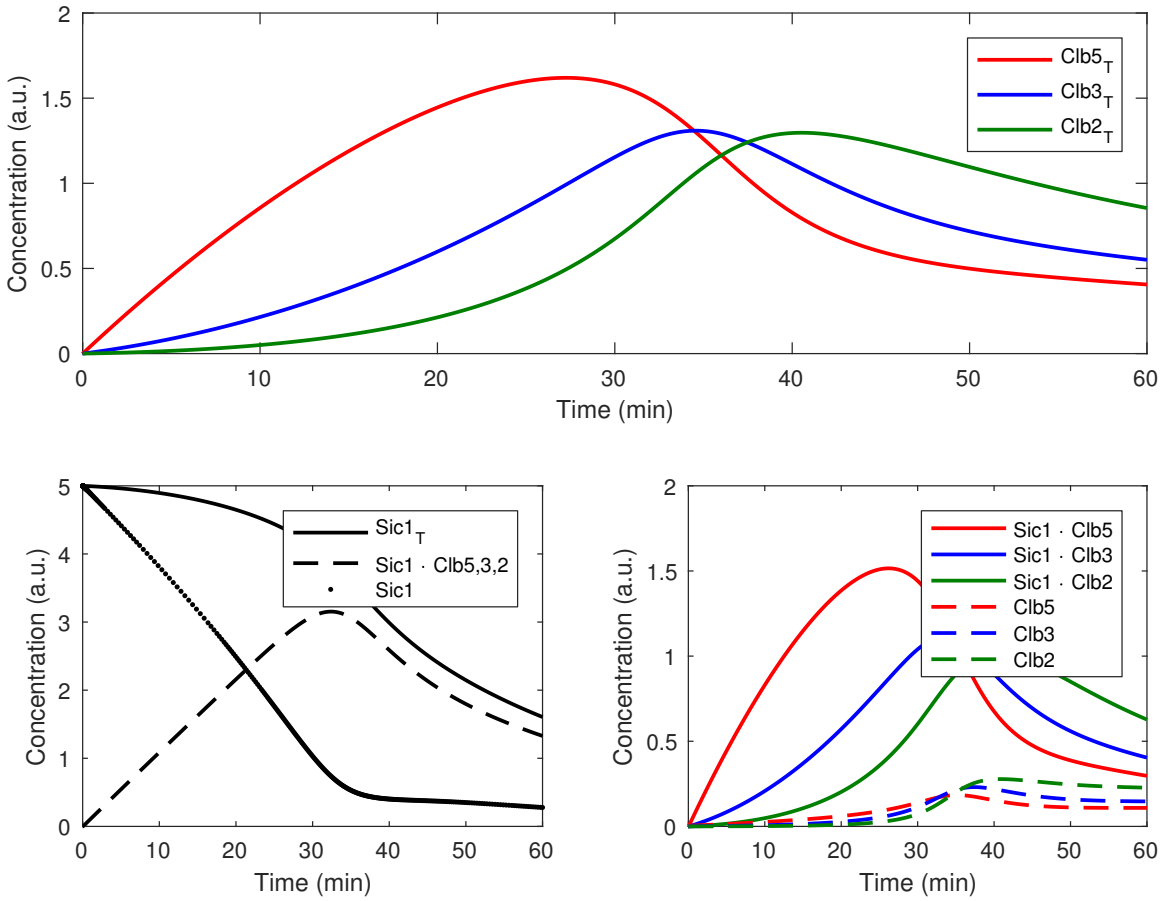


Figure S2.11: Time courses for *Design 2*. (Top) time courses for the total concentrations of the three Clb/Cdk1 complexes: Clb5/Cdk1, Clb3/Cdk1 and Clb2/Cdk1. (Bottom-left) time courses for total Sic1, Sic1 in complex with Clb/Cdk1 complexes, and Sic1. (Bottom-right) time courses for the binary (Clb/Cdk1) and ternary (Clb/Cdk1/Sic1) complexes.

To obtain sustained oscillations in *Design 2*, a number of parameters had to be altered. Ultimately, autonomous oscillations were obtained varying the following parameters as compared to the limit cycle parameter set for *Design 1A* and *Design 1B*: $v_x = 0.09$, $v_s = 0.2$, $\beta_s = 0.005$, $\delta = 0.05$. The resulting limit cycle is plotted in Fig. S2.12, and has a period of roughly 182 minutes.

The sensitivity analysis of the sustained oscillations is shown in Fig. S2.13. The results are generally similar to those obtained for *Design 1C.*

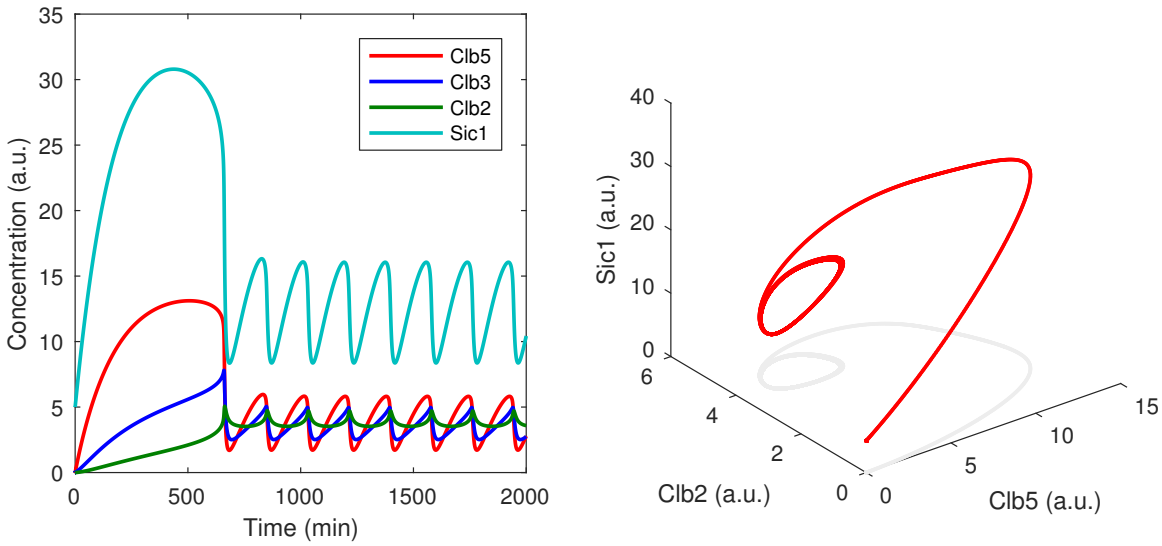


Figure S2.12: Sustained oscillations in *Design 2*. (Left) Limit cycle for the total concentrations of the four species. (Right) 3D view of the limit cycle in the Clb5-Clb2-Sic1 space.

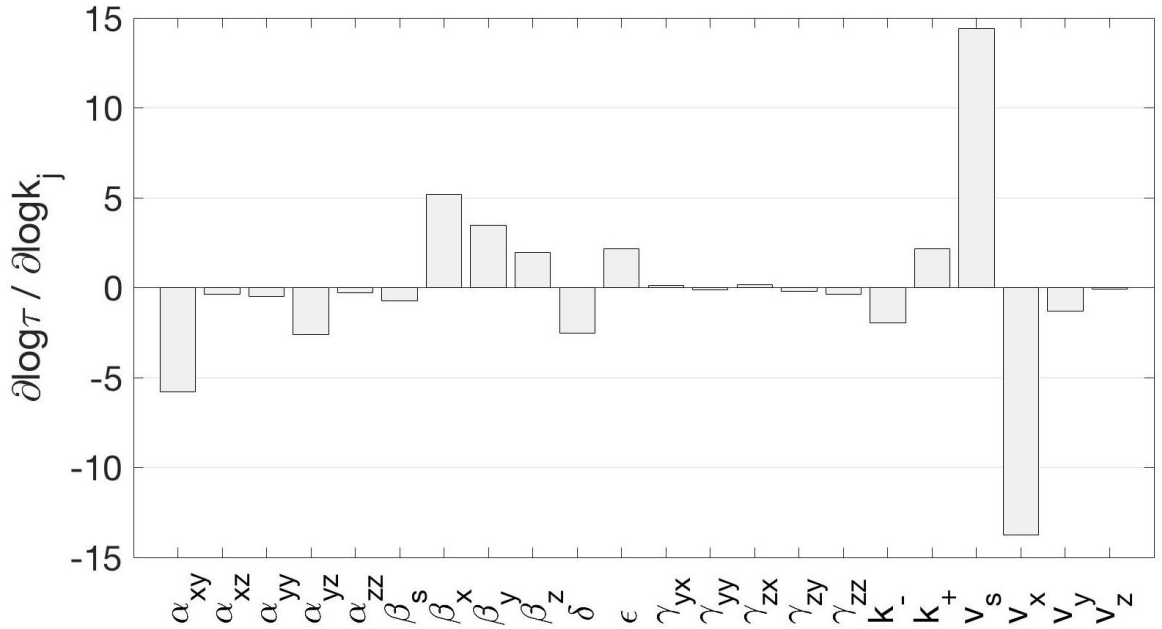


Figure S2.13: Logarithmic period derivatives for *Design 2*. Of note, despite some numeric changes, the general direction and relative size of the derivatives are similar as compared to the analysis of *Design 1C*.

Design 3: The *quasi-steady-state* (QSS) model

Design 3 refers to a model including a *quasi-steady-state* approximation on the ternary complex formation between Clb/Cdk1 complexes and Sic1 (Fig. S2.1C). Although this assumption involves removal of one parameter, the mathematical description of the model is significantly altered.

Equations for total concentrations

Starting from *Design 2*, the system of equations was re-written in terms of the total concentrations of the four species considered. The total concentration of the Clb/Cdk1 complexes, i.e. for Clb5: $x_T = [x] + [s \cdot x]$ and analogous formulas y_T and z_T, and Sic1, i.e. $[s_T] = [s] + [s \cdot x] + [s \cdot y] + [s \cdot z]$, evolve according to the following:

$$\begin{aligned}
\frac{d[x_T]}{dt} &= v_x - \beta_x[x] - \gamma_{yx}[x][y] - \gamma_{zx}[x][z] - \epsilon[s \cdot x] \\
\frac{d[y_T]}{dt} &= v_y - \beta_y[y] + \alpha_{xy}[x] + \alpha_{yy}[y] - \gamma_{zy}[z][y] - \gamma_{yy}[y]^2 - \epsilon[s \cdot y] \\
\frac{d[z_T]}{dt} &= v_z - \beta_z[z] + \alpha_{zz}[z] + \alpha_{xz}[x] + \alpha_{yz}[y] - \gamma_{zz}[z]^2 - \epsilon[s \cdot z] \\
\frac{d[s_T]}{dt} &= v_s - \beta_s[s] - (\epsilon + \delta([x] + [y] + [z]))\left([s \cdot x] + [s \cdot y] + [s \cdot z]\right). \qquad (2.4)
\end{aligned}$$

The *quasi-steady-state* approximation

We assume that Clb/Cdk1/Sic1 ternary complex formation and/or dissociation (k^+ and k^-) happen on a faster time-scale as compared to the other processes considered in the model. This implies that the ratio of active and ternary inactive complexes is at steady-state for a given Sic1 concentration. If we assume that ternary complex degradation (δ, ϵ) is relatively small, the equilibrium condition implies the following:

$$\begin{aligned}
k_-[x \cdot s] &= k_+[x][s] \\
[x \cdot s] &= K_A[x][s], \qquad (2.5)
\end{aligned}$$

with $K_A = \frac{k_+}{k_-} = \frac{1}{K_D}$, where K_A and K_D are the association and dissociation constants, respectively. Similar equations hold for y and z. This implies the following:

$$\begin{aligned}
[x_T] &= [x \cdot s] + [x] \\
&= [x]\left(1 + K_A[s]\right) \\
&= \frac{[x]}{f} \qquad (2.6)
\end{aligned}$$

where the variable $f[s]$ is introduced as follows:

$$f([s]) = \frac{1}{1 + [s]K_A}. \qquad (2.7)$$

Of note, f is a quantity that varies between 0 and 1. Since $[x] = f \cdot x_T$, f can be interpreted as the fraction of Clb/Cdk1 complex which is active, i.e. not in complex with Sic1, and that this fraction depends on the free Sic1 concentration

$[s]$. Therefore, $f[s]$ is also time-dependent since $[s]$ is time-dependent. $K_D = 1/K_A$ represents the concentration $[s]$ for which $f[s] = 1/2$. Of note, since $x = f \cdot [x_T]$ we have that $[x \cdot s] = (1 - f) \cdot [x_T]$.

The *quasi-steady-state* model

The quasi-steady-state assumption was incorporated in the expression of the total concentrations. The assumption can be considered for each Clb/Cdk1 species x, y or z independently. The introduced variable $f[s]$ depends on K_A, but K_A depends only on k^+ and k^-, which are the same for x, y and z. Hence, $f([s]$ applies to all three Clb/Cdk1 complexes.

The equation for f has to be supplemented by an equation specifying the free Sic1 concentration $[s]$ at any moment in time. From the mass-balance for s we derive that:

$$\begin{aligned}[s_T] &= [s] + [s \cdot x] + [s \cdot y] + [s \cdot z] \\ &= [s]\left(1 + K_A\left([x] + [y] + [z]\right)\right) \\ &= [s]\left(1 + K_A \frac{[x_T] + [y_T] + [z_T]}{1 + [s]K_A}\right). \end{aligned} \tag{2.8}$$

Multiplying out the fraction, we get a quadratic equation in $[s]$, as follows:

$$\begin{aligned}[s_T]\left(1 + [s]K_A\right) &= [s]\left(\left(1 + [s]K_A\right) + K_A\left([x_T] + [y_T] + [z_T]\right)\right) \\ 0 &= [s]^2 + \left(\frac{1}{K_A} + [x_T] + [y_T] + [z_T] - [s_T]\right)[s] - \frac{[s_T]}{K_A}. \end{aligned} \tag{2.9}$$

We find an expression for $[s]$ by solving quadratic equation and taking the positive root, as follows:

$$\begin{aligned}[s] = &-\frac{\frac{1}{K_A} + [x_T] + [y_T] + [z_T] - [s_T]}{2} \\ &+ \sqrt{\left(\frac{\frac{1}{K_A} + [x_T] + [y_T] + [z_T] - [s_T]}{2}\right)^2 + \frac{[s_T]}{K_A}}. \end{aligned} \tag{2.10}$$

Incorporating the *quasi-steady-state* assumption in the system for the total concentrations is achieved by writing down the equations for the total concentrations of the three Clb/Cdk1 species and Sic1 and replacing every occurrence of $[x]$ with $f[x_T]$ and $[s \cdot x]$ with $(1 - f)[x_T]$. This is done similarly for y and z. The complete QSS model is defined by a system of four first-order non-linear ODEs for the total concentrations of the species together with the equations for f and s. For clarity and brevity of the equations, the expressions for $[s]$ and f are written separately, but they are to be simply substituted into the ODEs. In summary:

$$
\begin{aligned}
[s] &= -\frac{\frac{1}{K_A} + [x_T] + [y_T] + [z_T] - [s_T]}{2} \\
&\quad + \sqrt{\left(\frac{\frac{1}{K_A} + [x_T] + [y_T] + [z_T] - [s_T]}{2}\right)^2 + \frac{[s_T]}{K_A}} \\
f &= \frac{1}{1 + [s]K_A} \\
\frac{d[x_T]}{dt} &= v_x - \beta_x f[x_T] - \epsilon(1-f)[x_T] - \gamma_{yx} f^2[x_T][y_T] - \gamma_{zx} f^2[x_T][z_T] \\
\frac{d[y_T]}{dt} &= v_y - \beta_y f[y_T] - \epsilon(1-f)[y_T] + \alpha_{xy} f[x_T] + \alpha_{yy} f[y_T] - \gamma_{yy} f^2[y_T]^2 \\
&\quad - \gamma_{zy} f^2[y_T][z_T] \\
\frac{d[z_T]}{dt} &= v_z - \beta_z f[z_T] - \epsilon(1-f)[z_T] + \alpha_{xz} f[x_T] + \alpha_{yz} f[y_T] + \alpha_{zz} f[z_T] \\
&\quad - \gamma_{zz} f^2[z_T]^2 \\
\frac{d[s_T]}{dt} &= v_s - \beta_s[s] - (\epsilon + \delta f([x_t] + [y_t] + [z_t]))(1-f)([x_T] + [y_T] + [z_T]).
\end{aligned}
\tag{2.11}
$$

The concentrations of the binary (Clb/Cdk1) and ternary (Clb/Cdk1/Sic1) complexes can be deduced by multiplying the total concentrations by f and $1-f$, respectively.

Numerical validation of the *quasi-steady-state* approximation

As a preliminary theoretical validation of our modeling approach, we explore the discrepancy between *Design 2* equations and the *quasi-steady-state* assumption in *Design 3*. If this assumption is satisfied, by high parameter values for Clb/Cdk1/Sic1 ternary complex formation and/or dissociation in *Design 2*, we would expect similar results between the two models. Conversely, if this assumption breaks down and these parameters become small, the two models should disagree in their behavior.

To test this assumption, we consider the canonical parameter set and simulate both models. We map the K_A parameter in the QSS model to the k^+ and k^- parameters in the standard model through the rule $k_+ = K_A k_-$. In the canonical parameter set $K_A = 10$. While keeping K_A constant (i.e. equal to 10), we can vary either k^+ or k^- and keep track of the model behavior. In Fig. S2.14, the discrepancy between the behavior of both models is shown, measured in terms of the Average Relative Difference, i.e. $ARD = \frac{1}{\#\text{species} \cdot \#\text{ time steps}} \sum_x \sum_t \left|\frac{x(t) - x^*(t)}{x(t)}\right|$ for the canonical parameter set but at different values for k_+ and k_- at constant ratio K_A, where x indicates the total concentrations of each of the four species (Sic1, Clb5, Clb3, Clb2) at time t in *Design 2*, and x* indicates the same in *Design 3*. t indicates the time points at which the ODE solver returns the solution. As

shown in Fig. S2.14, for high K_A values the QSS model numerically returns the same behavior as *Design 2*. Therefore, we expect all results to be roughly similar for both *Design 2* and *Design 3*.

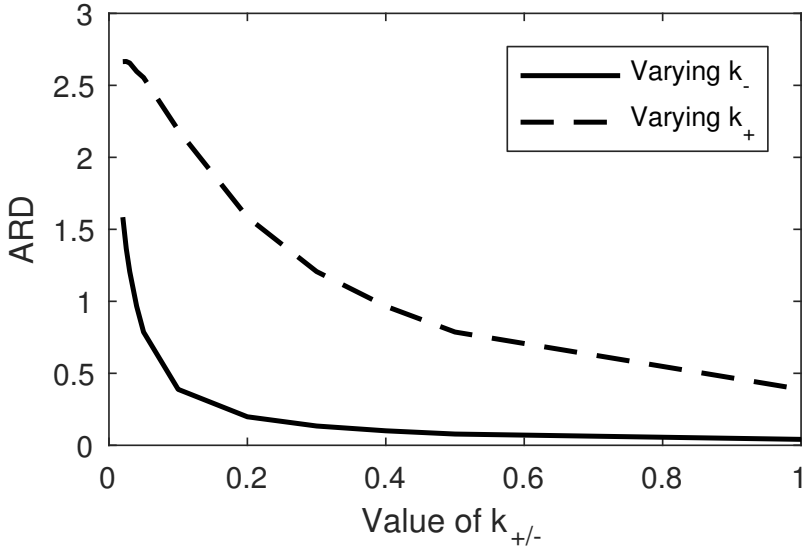


Figure S2.14: Analysis of the validity of the QSS assumption in the parameter space. The sum of squared errors (SSE) is measured between the model with (*Design 3*) and without (*Design 2*) the assumption. The k^- parameter is varied along the continuous line, whereas the k^+ parameter is varied along the dashed line, always for constant K_A. The two models exhibit a similar behavior for higher values of k^+ and k^-.

Transient oscillations for *Design 3* are plotted in Fig. S2.15. To obtain sustained oscillations in *Design 3*, we used the same parameter set as for *Design 2*. The resulting limit cycle is plotted in Fig. S2.16, and has a period of roughly 226 minutes.

The sensitivity analysis of the sustained oscillations is shown in Fig. S2.17. The results are generally similar to those obtained for *Design 2* and *Design 1C*.

Summary of autonomous oscillations for designs 1A–3

All five models corresponding to designs 1A-3 (illustrated in Fig. S2.1) are able to yield sustained oscillations in the form of limit cycles. For each design, initial parameter sets resulting in limit cycles were identified through a manual adjustment of one or several model parameters through the slider functionality in COPASI [73], starting from the canonical parameter set [10] (Table S2.1).

We quantified the control on the period of the model parameters (Fig. S2.4, S2.7, S2.10, S2.13 and S2.17). The analysis of our models indicates that this control is shared among several parameters, as it has been observed in other biochemical oscillatory systems for which the function of the oscillation is unclear [83]. This is the first time that a distributed control is found for an oscillation as functional as an autonomous cell cycle. The results recover the fact that the sum of the con-

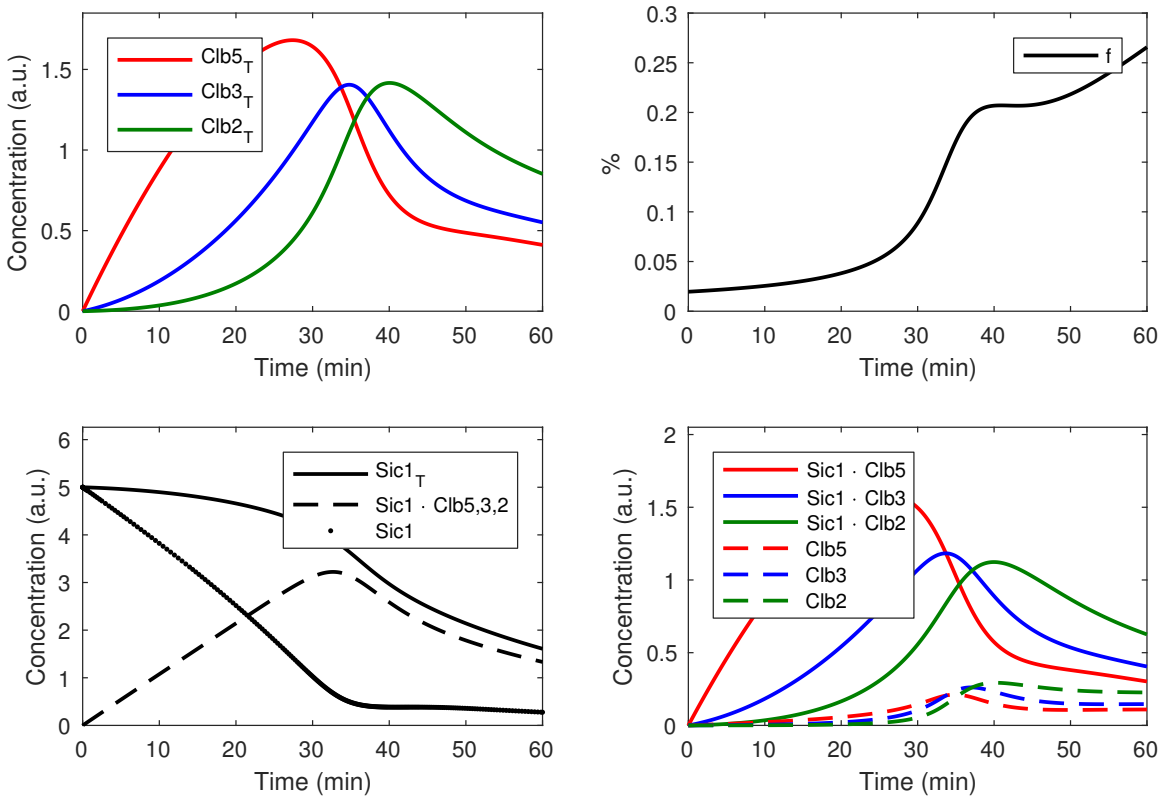


Figure S2.15: Time courses for *Design 3*. (Top-left) time courses for the total concentrations of the three Clb/Cdk1 complexes: Clb5/Cdk1, Clb3/Cdk1 and Clb2/Cdk1. (Top-tight) Plot of the function f over time. (Bottom-left) time courses for total Sic1, Sic1 in complex with Clb/Cdk1 complexes, and Sic1. (Bottom-right) time courses for the binary (Clb/Cdk1) and ternary (Clb/Cdk1/Sic1) complexes.

trol coefficients on the period of oscillations for all the parameters with dimension 1/time must equal –1 [77]. Control coefficients of (–)1 indicate that the controlled property is (inversely) proportional to the controlling parameter. Among the parameters that exerted a more considerable control in Design 3 (i.e. an absolute value of the control coefficient > 2), some increased (v_s, β_x, β_y, K_A, ϵ) and some decreased (negative control coefficient; v_x, α_{xy}, α_{yz}, δ) the period of oscillations. The latter correspond to the parameters that activate cell cycle progression by decreasing the period. Indeed, they stimulate the formation of Clb5/Cdk1, Clb3/Cdk1, Clb2/Cdk1 and the degradation of Sic1, respectively. Conversely, activation of the former should inhibit cell cycle progression, by stimulating Sic1 formation, dissociation of Clb5/Cdk1 and Clb3/Cdk1 and by favoring the formation and degradation of the ternary Clb/Cdk1/Sic1 complex. The control that these parameters exhibit is conserved between designs 1C, 2 and 3 (Fig. S2.10, Fig. S2.13 and Fig. S2.17, respectively). Strikingly, the Clb3 PFL (associated to the parameter α_{yy}), the Clb3 NFL (γ_{yy}) – as well as the Clb2 PFL (α_{zz}) and the Clb2 NFL (γ_{zz}) – and the four known Clb-regulated inhibitory regulations mediated by both the Clb/Cdk1 complexes and Anaphase-Promoting Complex, APC (see

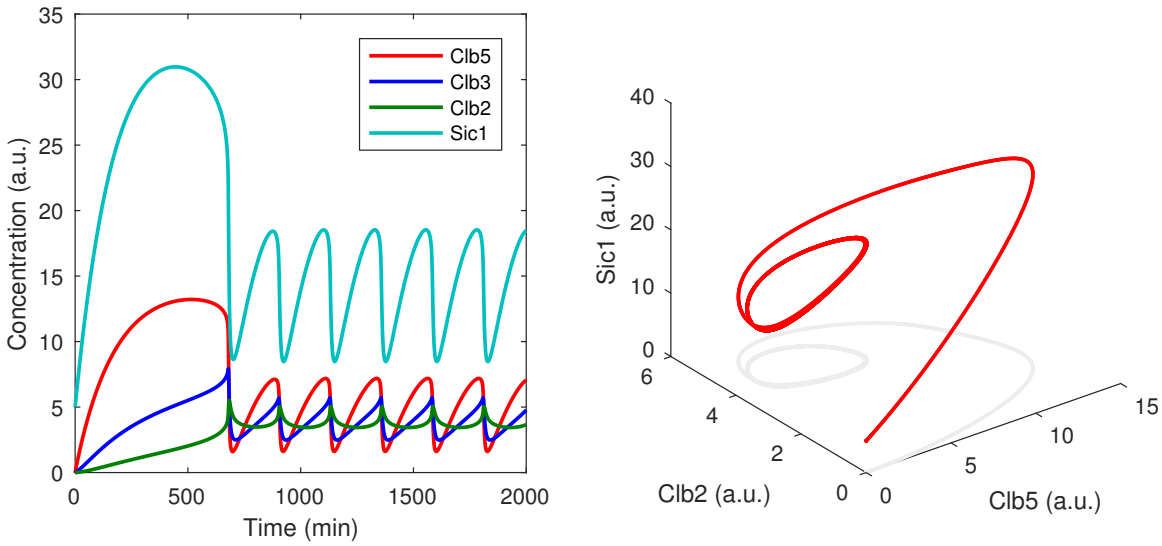


Figure S2.16: Sustained oscillations in *Design 3.* (Left) Limit cycle for the total concentrations of the four species. (Right) 3D view of the limit cycle in the Clb5-Clb2-Sic1 space.

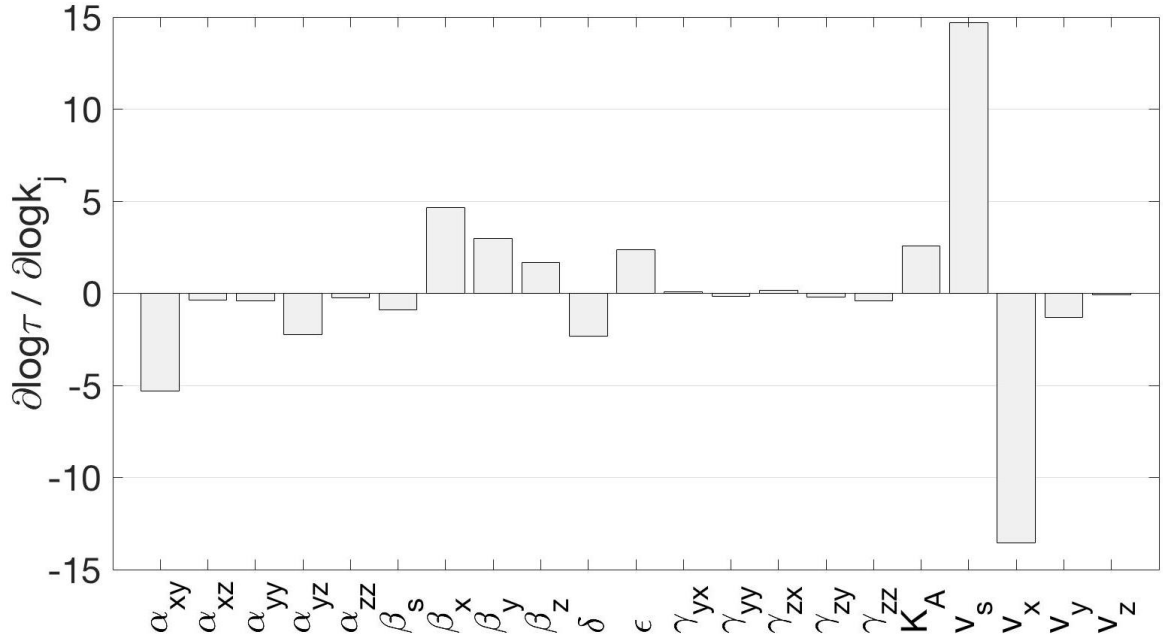


Figure S2.17: Logarithmic period derivatives for *Design 3.* Of note, despite some numeric changes, the general direction and relative size of the derivatives are similar as compared to the analysis of *Design 2*.

Section 2.4) exerted almost no control over the period of oscillations.

Designs 4 through *9*: Inhibitory regulations that may boost oscillatory potential

Below we briefly highlight the changes in the equations for *Design 4* through *Design 9* as compared to *Design 3*.

Design 4

Design 4 entails inhibition of Clb5 synthesis by Clb2/Cdk1. The SBF transcription factor, formed by Swi4 and Swi6, promotes transcription of the G1 phase cyclin genes *CLN1* and *CLN2*; the MBF transcription factor, formed by Mbp1 and Swi6, promotes transcription of the S phase cyclin genes *CLB5* and *CLB6*. Genetic evidence indicates that *CLB2* and *SWI6* are functionally related [84], and Clb2 has been shown to interact physically with Swi4, thus repressing transcription of the G1 cyclins [44]. Inhibition of the G1 cyclins translates to an effective inhibition of the Clb5/Cdk1 activity, due to the lack of the PFL between Cln2/Cdk1 and SBF/MBF [45] and to the lifted inhibition of Sic1 by Cln1,2/Cdk1 [46].

The altered ODE for Clb5 (x) is now written as follows:

$$\frac{d[x_T]}{dt} = \frac{v_x}{1+\frac{[z_T]}{K_{zx}}} - \beta_x f[x_T] - \epsilon(1-f)[x_T] - \gamma_{yx} f^2[x_T][y_T] - \gamma_{zx} f^2[x_T][z_T] \quad (2.12)$$

This structural change in the equations with the inhibitory term is representative for all subsequent designs. Of note, v_x, which used to represent the synthesis of Clb5, is now the V_{max} of the synthesis, which is attained when Clb2 is not present.

Design 5

Design 5 entails inhibition of Sic1 synthesis by Clb2/Cdk1 through the *SWI5* transcription factor. During the G2 phase, Cdk1 phosphorylates specific serine residues of Swi5 near the NLS (Nuclear Localization Sequence) at its C-terminal, in order to keep Swi5 sequestered in the cytoplasm [47], effectively inhibiting *SIC1* transcription. The Cdk-dependent phosphorylation reaction may be reversed by the phosphatase Cdc14 [85], which thus contributes to *SIC1* transcription. *Design 5* describes the inhibition of *SIC1* transcription mediated by the Clb/Cdk1 activity, reflecting the likely scenario where the most abundant Cdk1 activity is due to Clb2/Cdk1.

The altered ODE for Sic1 (s) is now written as follows:

$$\frac{d[s_T]}{dt} = \frac{v_s}{1+\frac{[z_T]}{K_{zs}}} - \beta_s[s] - (\epsilon + \delta f([x_t]+[y_t]+[z_t]))(1-f)([x_T]+[y_T]+[z_T]) \quad (2.13)$$

Design 6

Design 6 entails inhibition of Sic1 synthesis by Clb2/Cdk1, Clb3/Cdk1 and Clb5/Cdk1 through the *SWI5* transcription factor. *Design 6* describes the same

mechanism detailed for *Design 5* but mediated by the three Clb/Cdk1 complexes: Clb2/Cdk1, Clb3/Cdk1 and Clb5/Cdk1.

The altered ODE for Sic1 (s) is now written as follows:

$$\begin{aligned}\frac{d[s_T]}{dt} = & \frac{v_s}{1+\frac{[x_T]+[y_T]+[z_T]}{K_{cs}}} - \beta_s[s] \\ & - (\epsilon + \delta f([x_t]+[y_t]+[z_t]))(1-f)([x_T]+[y_T]+[z_T])\end{aligned} \tag{2.14}$$

Design 7

Design 7 entails inhibition of Clb2 and Clb3 syntheses by Sic1. A recent study that integrated experimentation and computer modeling showed that Sic1 oscillations rescue viability of cells with low levels of mitotic Clb cyclins [48]. However, the molecular mechanism(s) at the basis of this observation at the moment remains obscure. Here we propose that inhibition of Clb2 and Clb3 synthesis by Sic1 may rationalize this observation; both *CLB2* and *CLB3* genes appear to be regulated by a similar transcriptional mechanism [32], thus we have incorporated for both the same Sic1-mediated inhibitory regulation.

The altered ODEs for Clb3/Cdk1 (y) and Clb2/Cdk1 (z) are now written as follows:

$$\begin{aligned}\frac{d[y_T]}{dt} = & \frac{v_y}{1+\frac{[s_T]}{K_{syz}}} - \beta_y f[y_T] - \epsilon(1-f)[y_T] + \alpha_{xy} f[x_T] + \alpha_{yy} f[y_T] \\ & - \gamma_{yy} f^2 [y_T]^2 \\ \frac{d[z_T]}{dt} = & \frac{v_z}{1+\frac{[s_T]}{K_{syz}}} - \beta_z f[z_T] - \epsilon(1-f)[z_T] + \alpha_{xz} f[x_T] + \alpha_{yz} f[y_T] + \alpha_{zz} f[z_T] \\ & - \gamma_{zz} f^2 [z_T]^2\end{aligned} \tag{2.15}$$

Of note, *Design 7* is slightly different from *Design 4*, *Design 5* and *Design 6*, as the inhibitory term is assumed to affect basal synthesis while there are other synthesis terms in the equations that are not affected.

Design 8

Design 8 entails inhibition of Clb2, Clb3 and Clb5 syntheses by Sic1. The altered ODEs for Clb5/Cdk1, Clb3/Cdk1 and Clb2/Cdk1 are now written as follows:

$$\begin{aligned}
\frac{d[x_T]}{dt} &= \frac{v_x}{1+\frac{[s_T]}{K_{sxyz}}} - \beta_x f[x_T] - \epsilon(1-f)[x_T] - \gamma_{yx} f^2[x_T][y_T] \\
&\quad - \gamma_{zx} f^2[x_T][z_T] \\
\frac{d[y_T]}{dt} &= \frac{v_y}{1+\frac{[s_T]}{K_{sxyz}}} - \beta_y f[y_T] - \epsilon(1-f)[y_T] + \alpha_{xy} f[x_T] + \alpha_{yy} f[y_T] \\
&\quad - \gamma_{yy} f^2[y_T]^2 \\
&\quad - \gamma_{zy} f^2[y_T][z_T] \\
\frac{d[z_T]}{dt} &= \frac{v_z}{1+\frac{[s_T]}{K_{sxyz}}} - \beta_z f[z_T] - \epsilon(1-f)[z_T] + \alpha_{xz} f[x_T] + \alpha_{yz} f[y_T] + \alpha_{zz} f[z_T] \\
&\quad - \gamma_{zz} f^2[z_T]^2
\end{aligned} \tag{2.16}$$

Design 9

Design 9 entails inhibition of Sic1 synthesis by Sic1 through the *SWI5* transcription factor. This regulation does not have any experimental support; however, we aimed to test all possible Sic1-dependent negative regulations.

The altered ODE for Sic1 is now written as follows:

$$\frac{d[s_T]}{dt} = \frac{v_s}{1+\frac{[s_T]}{K_{ss}}} - \beta_s[s] - \left(\epsilon + \delta f\left([x_t]+[y_t]+[z_t]\right)\right)(1-f)\left([x_T]+[y_T]+[z_T]\right) \tag{2.17}$$

System Design Space (SDS) methodology

In this work we make an extensive use of the System Design Space (SDS) methodology developed by Savageau and collaborators [25–27, 62] and the associated Python toolbox [31]. The cited papers should be seen as required reading to reproduce the analyses presented in our work. We especially point the reader to the Supplementary Information accompanying [27], which contains an excellent introduction to the terminology with clear examples. The SDS methodology has been proposed to overcome the analytical difficulty to find limit cycles. In 1900, David Hilbert posed his 16$^{\text{th}}$ problem, which partially concerns the finding of the number of limit cycles of a polynomial differential equation in the plane [28]. The Bendixson-Dulac theorem and the Poincaré-Bendixson theorem predict the absence or existence, respectively, of limit cycles of two-dimensional nonlinear dynamical systems. However, these theorems do not help to actually to find the parameter sets that generate limit cycles.

The SDS methodology has grown out of the Biochemical Systems Theory (BST), in which every process (generalized mass-action reaction [86] is formulated in a simplified way as a product of power-law functions [87]. A trade-off exists between (i) the accuracy by which system's complexity is modeled and (ii) the computational cost to analyze a model and the complexity of its results. The SDS methodology starts from a system of ordinary differential equations (ODEs), described by using generalized mass-action (GMA) kinetics [87]. First, the set of all combinatorically possible combinations of single dominant positive and negative terms in each ODE is generated. The reduction to dominant processes transforms the ODE system into an S-system. S-systems can capture Saturable and Synergistic (thus the capital S) properties of a biochemical system [88] and is then referred to as a *phenotype* that can be used to approximate the full GMA model [25]. In an S-system, for a particular phenotype, every ODE consists of a single dominant positive term and a single dominant negative term. For a given set of parameters and concentrations there exists a single dominant positive term, i.e. largest, and a single dominant negative term in each differential equation. The dominance of certain positive and negative terms gives rise to *dominance* conditions, i.e. inequalities stating that the dominant positive (negative) term is larger than the other positive terms in a specific ODE. Altogether, the dominance conditions form a set of inequalities that are either inconsistent, i.e. there is no set of parameters and concentrations that satisfies them all, or consistent. When reducing the mathematical description of the phenotypes to these dominant processes, a biochemical system becomes mathematically tractable, and a consistent set of dominance conditions defines boundaries within the parameter space and (reaction) state space within which the dominance conditions, and therefore the phenotype, are valid [27, 88]. In this way, a phenotype may be viewed as a bounded area within the parameter and state space. Subsequently, by transforming the equations to logarithmic coordinates, the S-system becomes linear, and an analytical solution for the steady states may be obtained [25]. In addition to this analytical solution, properties such as the stability of steady states may be determined. This is particularly relevant for the identification of limit cycles, since Hopf bifurcations that give rise to limit cycles occur when a pair of complex conjugate eigenvalues crosses the imaginary axis. Consequently, a fixed point (steady state in the mathematical term) with two complex conjugate eigenvalues with positive real parts is a necessary condition for the occurrence of limit cycles after undergoing a Hopf bifurcation. Therefore, by using the SDS methodology, phenotypes with unstable steady states that have two complex conjugate eigenvalues with positive real parts may suggest bounded areas of the parameter space that might generate limit cycles, as highlighted previously [27]. This procedure greatly reduces the area of the parameter space to be sampled, and allows for the exploration of relatively small areas that might otherwise be overlooked.

SDS methodology and GMA casting application to *Design 7*

We start here by deriving the 11 different model designs that we considered in our work. For designs *1A*, *1B*, *1C*, *2* and *3*, we show (i) transient oscillations

using a canonical parameter set (see Section 2.4, Table S2.1), an initial parameter set yielding limit cycles, and (ii) a sensitivity analysis for all parameters on the period of oscillations. *Design 4* through *Design 9* develop on *Design 3*, and for these we highlight the changes that occur in the equations.

To implement any design that we considered in our work such that the models work with the System Design Space Toolbox, these need to be translated into their GMA (Generalized Mass Action) form. In the following, we use *Design 7* to illustrate the translation from the Ordinary Differential Equations (ODEs) into the Generalized Mass-Action (GMA) form. In the GMA form, all equations must consist of sums of products of parameters and concentrations that may be raised to a power.

For *Design 7*, the ODEs are written as follows:

$$
\begin{aligned}
\frac{d[x_T]}{dt} &= v_x - \beta_x f[x_T] - \epsilon(1-f)[x_T] - \gamma_{yx} f^2[x_T][y_T] - \gamma_{zx} f^2[x_T][z_T] \\
\frac{d[y_T]}{dt} &= \frac{v_y}{1+\frac{[s_T]}{K_{syz}}} - \beta_y f[y_T] - \epsilon(1-f)[y_T] + \alpha_{xy} f[x_T] + \alpha_{yy} f[y_T] - \gamma_{yy} f^2[y_T]^2 \\
&\quad - \gamma_{zy} f^2[y_T][z_T] \\
\frac{d[z_T]}{dt} &= \frac{v_z}{1+\frac{[s_T]}{K_{syz}}} - \beta_z f[z_T] - \epsilon(1-f)[z_T] + \alpha_{xz} f[x_T] + \alpha_{yz} f[y_T] + \alpha_{zz} f[z_T] \\
&\quad - \gamma_{zz} f^2[z_T]^2 \\
\frac{d[s_T]}{dt} &= v_s - \beta_s[s] - \left(\epsilon + \delta f\left([x_t]+[y_t]+[z_t]\right)\right)(1-f)\left([x_T]+[y_T]+[z_T]\right). \\
f([s]) &= \frac{1}{1+[s]K_A} \\
[s] &= -\frac{\frac{1}{K_A}+[x_T]+[y_T]+[z_T]-[s_T]}{2} \\
&\quad + \sqrt{\left(\frac{\frac{1}{K_A}+[x_T]+[y_T]+[z_T]-[s_T]}{2}\right)^2 + \frac{[s_T]}{K_A}}
\end{aligned}
\tag{2.18}
$$

For the GMA form, we need to get rid of any fractions, i.e. in the equations for f and s, and we need to expand brackets, i.e. in the equations for x, y, z and s. For the $(1-f)$ terms, we introduce a new variable f_{inv}; for the $[x_T]+[y_T]+[z_T]$ terms, we introduce a new variable $\mathrm{Clb_T}$. In the equations for f and $\mathrm{s_{free}}$, we introduce several auxiliary variables to satisfy the GMA form. Ultimately, the

GMA form is written as follows:

$$
\begin{aligned}
\frac{d[x_T]}{dt} &= v_x - \beta_x f[x_T] - \epsilon f_{inv}[x_T] - \gamma_{yx} f^2[x_T][y_T] - \gamma_{zx} f^2[x_T][z_T] \\
\frac{d[y_T]}{dt} &= v_y \, \text{aux}_3^{-1} - \beta_y f[y_T] - \epsilon f_{inv}[y_T] + \alpha_{xy} f[x_T] + \alpha_{yy} f[y_T] - \gamma_{yy} f^2[y_T]^2 \\
&\quad - \gamma_{zy} f^2[y_T][z_T] \\
\frac{d[z_T]}{dt} &= v_z \, \text{aux}_3^{-1} - \beta_z f[z_T] - \epsilon f_{inv}[z_T] + \alpha_{xz} f[x_T] + \alpha_{yz} f[y_T] + \alpha_{zz} f[z_T] \\
&\quad - \gamma_{zz} f^2[z_T]^2 \\
\frac{d[s_T]}{dt} &= v_s - \beta_s [s_{free}] - (\epsilon + \delta f \text{CLB}_T) \, f_{inv} \text{CLB}_T \\
f([s]) &= \frac{1}{f_{\text{denom}}} \\
f_{\text{denom}} &= 1 + [s] K_A \\
f_{inv} &= 1 - f \\
\text{CLB}_T &= [x_T] + [y_T] + [z_T] \\
[s] &= -\frac{1}{2} K_A^{-1} + \frac{1}{2} + [s_T] - \frac{1}{2} \text{CLB}_T + \text{aux}_1^{\frac{1}{2}} \\
\text{aux}_1 &= [s_T] K_A^{-1} - \frac{1}{4} \text{aux}_2^2 \\
\text{aux}_2 &= K_A^{-1} - [s] + \text{CLB}_T \\
\text{aux}_3 &= 1 + [s_T] K_{syz}^{-1}
\end{aligned}
\tag{2.19}
$$

In Python code that is readable by the System Design Space Toolbox, the equations are written as follows:

```
Eq = [
'x. = v_x - b_x * f * x - g_yx * f^2 * x * y - g_zx * f^2 * x * z
     - e * f_inv * x',
'y. = v_y * aux3^(-1.0) - b_y * f * y + a_xy * f * x + a_yy * f *
     y - g_yy * f^2 * y^2 - g_zy * f^2 * z * y - e * f_inv * y',
'z. = v_z * aux3^(-1.0) - b_z * f * z + a_xz * f * x + a_yz * f *
     y + a_zz * f * z - g_zz * f^2 * z^2 - e * f_inv * z',
's. = v_s - b_s * s_free - e * f_inv * clbT - d * f_inv * clbT *
    f * clbT',
's_free = -(1/2.0)*K_A^(-1.0) + (1/2.0)*s - (1/2.0)*clbT + aux1
    ^(1/2.0)',
'f_inv = 1 - f',
'f = f_denom^(-1.0)',
'f_denom = (1+s_free*K_A)',
'aux1 = s*K_A^(-1.0) + (1/4.0)*aux2^2.0',
'aux2 = K_A^(-1.0) - s + clbT',
'clbT = x+y+z',
'aux3 = 1 + s*K_syz^(-1.0)']
```

Interaction	Symbol	References	Notes
Sic1 $\leftrightarrow$ Clb5,3,2	$k^{+/-}$	[10]	Fast complex formation is incorporated in Design 3-9 (Mart Loog, personal communication)
Clb5 $\rightarrow$ Clb3	α_{xy}	[32], [33, 89–91]	
Clb5 $\rightarrow$ Clb2	α_{xz}	[33, 89–91], [32, 92, 93]	
Clb3 $\rightarrow$ Clb3	α_{yy}	[32], [32]*	
Clb3 $\rightarrow$ Clb2	α_{yz}	[32, 92, 93] , [32]*	
Clb2 $\rightarrow$ Clb2	α_{zz}	[33, 89–91], [32, 92, 93]	
Clb2 $\dashv$ Clb2	γ_{zy}	[94], [82]	Absent in Design 1A
Clb2 $\dashv$ Clb3	γ_{zy}	[10], [94],	Absent in Design 1A. We hypothesize that Clb3 is targeted for degradation by APC-Cdc20
Clb2 $\dashv$ Clb5	γ_{zx}	[94], [95, 96]	Absent in Design 1
Clb3 $\dashv$ Clb3	γ_{yy}	[10]	Absent in Design 1A. We hypothesize that Clb3 may activate APC-Cdc20. We hypothesize that Clb3 is targeted for degradation by APC-Cdc20
Clb3 $\dashv$ Clb5	γ_{yx}	[10], [95, 96]	Absent in Design 1A. We hypothesize that Clb3 may activate APC-Cdc20
Clb5,3,2 $\rightarrow$ Sic1-Clb5,3,2	δ	[46, 97]	
Clb2 $\dashv$ Clb5	K_{zx}	[44]	Present only in Design 4
Clb2 $\dashv$ Sic1	K_{zs}	[47]	Present only in Design 5-6
Clb3,5 $\dashv$ Sic1	K_{cs}	–	Present only in Design 6. We hypothesize that Clb3/Cdk1 and Clb5/Cdk1 may inhibit SIC1 transcription
Sic1 $\dashv$ Clb2,3	K_{syz}	[48]	Present only in Design 7-8. We hypothesize that Sic1 may inhibit CLB2\CLB3 transcription
Sic1 $\dashv$ Clb5	K_{sxyz}	[48]	Present only in Design 8. We hypothesize that Sic1 may inhibit CLB5 transcription
Sic1 $\dashv$ Sic1	K_{ss}	–	Present only in Design 9. We hypothesize that Sic1 may inhibit SIC1 transcription

Table S2.2: Overview of the experimental evidence for interactions and regulations in model designs *1A-9*, along with the associated parameter symbol and relevant notes. The symbol -> indicates activations, $\dashv$ indicates inhibitions, and $\leftrightarrow$ indicates reversible complex formation. For clarity, the relevant references are grouped based on the described interaction/regulation, see Table S2.3.

Experimental evidence of the interactions and regulations in designs 1–9

The known experimental evidence for the regulations across all designs used in this work are listed in Table S2.2 and S2.3.

Reference	Summary of results used in Table S2.2
[10]	Sic1 interacts and co-exists in time with Clb5/Cdk1, Clb3/Cdk1 and Clb2/Cdk1.
[32]	Fkh2 regulates CLB3 expression.
[32]*	Clb3,4/Cdk1 play a role in Fkh2 phosphorylation.
[33, 89–91]	Clb5/Cdk1 and Clb2/Cdk1 interact with, and phosphorylate, Fkh2 to control Clb1,2 accumulation.
[32, 92, 93]	CLB1,2 transcription is regulated by Fkh2 during the G2/M phase.
[94]	Phosphorylation of Cdc20 by Clb2/Cdk1 activates APC-Cdc20.
[82]	APC-Cdc20 degrades mitotic cyclins.
[95, 96]	APC-Cdc20 targets Clb5 for degradation.
[46, 97]	Clbs/Cdk1 phosphorylate Sic1, resulting in the recognition of Sic1 by the protein degradation machinery.
[44]	Clb2/Cdk1 interacts with Swi4 and represses G1 cyclins transcription. This regulation translates to an inhibition of Clb5/Cdk1, due to the lack of the PFL between Cln2/Cdk1 and SBF/MBF [45] and to the lifted inhibition of Sic1 by Cln1,2/Cdk1 [46, 97].
[47]	Inhibition of SIC1 transcription is mediated by Clb2/Cdk1.
[48]	Hypothetical interaction to rationalize the observation that Sic1 oscillations rescue viability of cells with low levels of mitotic cyclins (Clb2).

Table S2.3: Summary of the results represented by the grouped references from Table S2.2.

Parameter correlation analysis in limit cycles

Within the set of identified limit cycles for each model design, we analyzed whether there were correlations present between the parameters as measured by the Pearson correlation coefficient. Figures summarizing the correlation coefficients as heatmaps between all model parameters across the seven model designs that returned limit cycles are available through the Supplementary Code Repository. In Table S2.4, the set of combinations of two parameters that were highly correlated (absolute value of correlation coefficient ≥ 0.5) in three or more of the designs are summarized.

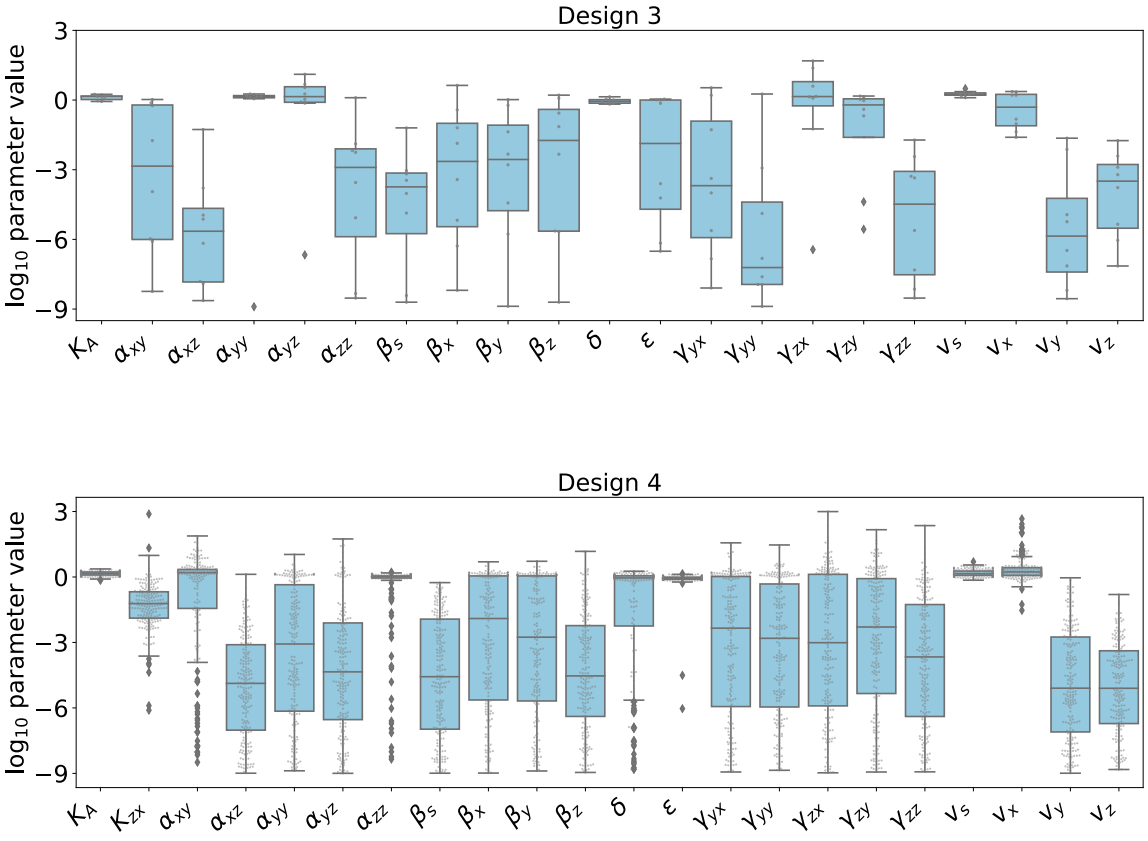
Spread of limit cycles across the parameter space

To indicate the unlikeliness of the connectivity between the identified parameter sets that result in limit cycles, boxplots of the parameter values in these parameter sets for designs *3–9* were generated (Fig. S2.18). For all designs most parameter values cover multiple orders of magnitude. We do note that the logarithmic scales

Parameter #1	Parameter #2	Designs	# Pos. correlations	# Neg. correlations
α_{yy}	γ_{zy}	3, 4, 5, 6, 7, 9	6	0
α_{xy}	α_{xz}	3, 5, 8, 9	4	0
v_s	v_x	3, 6, 7, 8	4	0
K_A	v_s	3, 4, 6, 8	0	4
γ_{zx}	v_x	6, 7, 8, 9	4	0
α_{xy}	β_y	3, 5, 6, 7	4	0
α_{xz}	β_z	3, 5, 7	3	0
β_z	ϵ	3, 7, 9	0	3

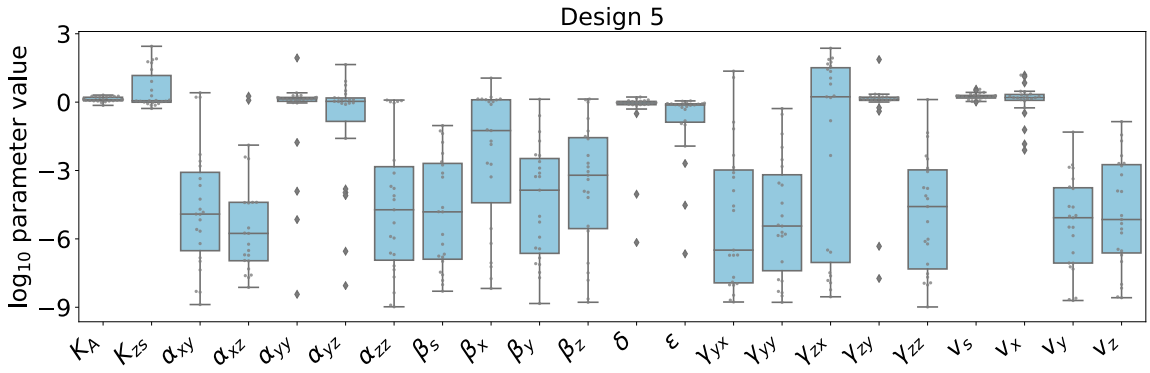
Table S2.4: Combinations of two parameters that were highly correlated (absolute value of correlation coefficient ≥ 0.5) in three or more of the model designs.

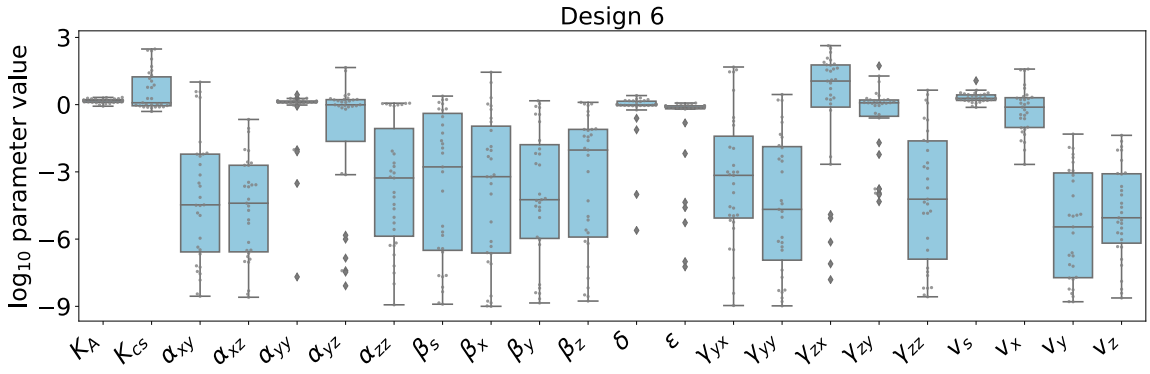
on the lower end, e.g. -3 through -9, should not be given too much weight, since this may simply indicate that the parameter value is so small that would not affect the model output. However, even discounting the lower range, the parameter values in the limit cycles still cover a wide range of values.

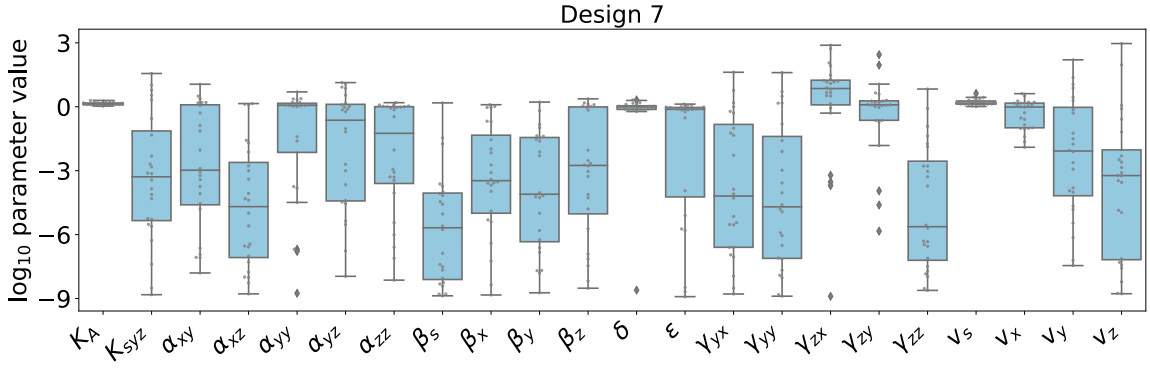


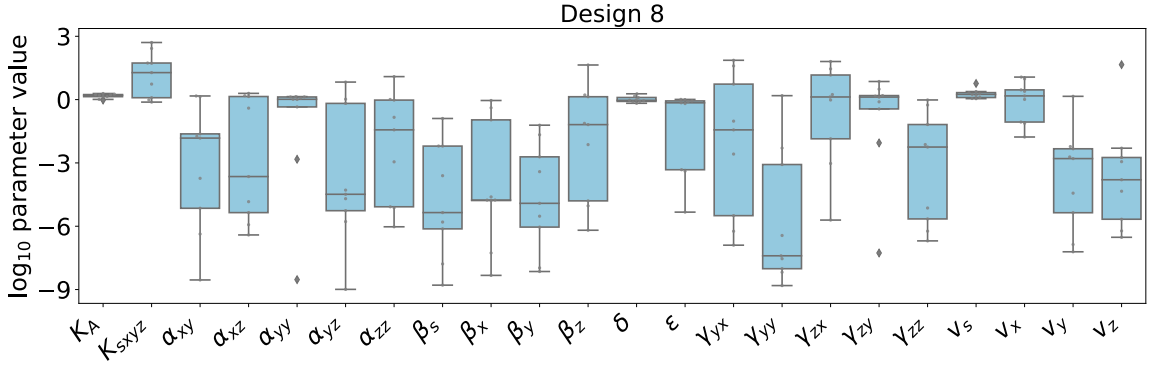
Limit cycles in designs *3–9* differ in period and amplitude

Fig. S2.19 and S2.19 display the periods and amplitudes (defined as the maximum ratio across the four species of the minimum and maximum of the oscillation) retrieved for the limit cycles of Design 1-9 respectively.


Design 5
log10 parameter value
3
0
−3
−6
−9
KA
Kzs
αxy
αxz
αyy
αyz
αzz
βs
βx
βy
βz
δ
ε
γyx
γyy
γzx
γzy
γzz
νs
νx
νy
νz


Design 6
log10 parameter value
3
0
−3
−6
−9
KA
Kcs
αxy
αxz
αyy
αyz
αzz
βs
βx
βy
βz
δ
ε
γyx
γyy
γzx
γzy
γzz
νs
νx
νy
νz


Design 7
log10 parameter value
3
0
−3
−6
−9
KA
Ksyz
αxy
αxz
αyy
αyz
αzz
βs
βx
βy
βz
δ
ε
γyx
γyy
γzx
γzy
γzz
νs
νx
νy
νz


Design 8
log10 parameter value
3
0
−3
−6
−9
KA
Ksxyz
αxy
αxz
αyy
αyz
αzz
βs
βx
βy
βz
δ
ε
γyx
γyy
γzx
γzy
γzz
νs
νx
νy
νz

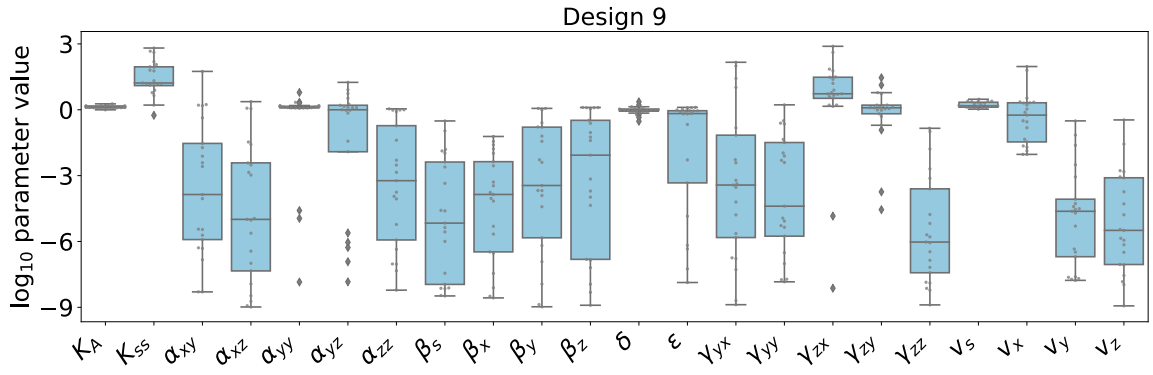


Figure S2.18: Boxplots of the parameter values in the limit cycles identified for designs *3–9* on log10 scale. The entire allowable range in our parameter sampling $[10^3, 10^{-9}]$ is shown. The box edges indicate the 25th and 75th percentile respectively and the black line within the box indicates the median of the parameter values. The actually sampled parameter values are indicates by grey dots.

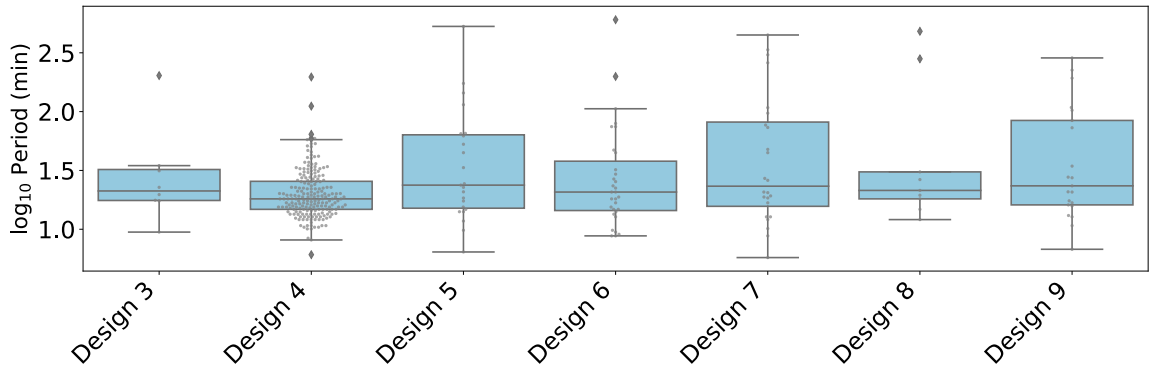


Figure S2.19: Boxplots of the logarithm of the period of the limit cycles identified for designs *3–9*. The periods roughly cover the interval $[10, 1000]$ minutes.

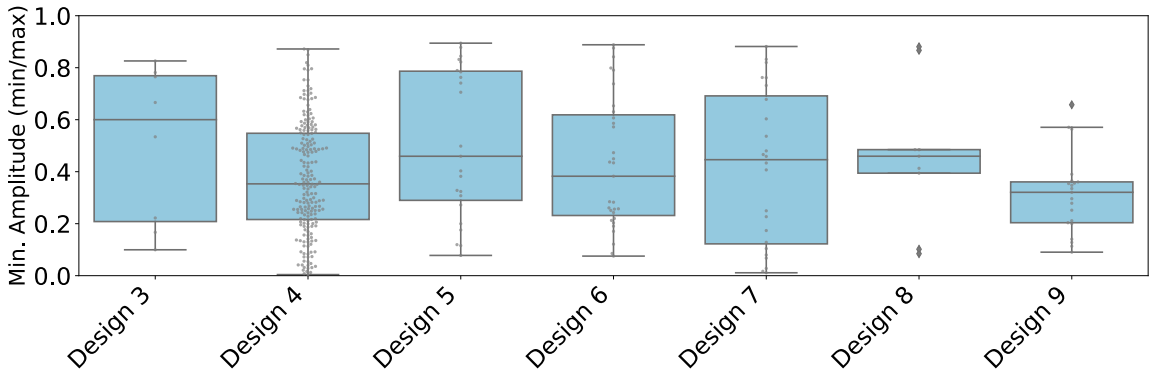


Figure S2.20: Boxplots of the amplitude defined as the maximum across the four species of the minimum/maximum ratio of the limit cycle time courses designs *3–9*. We required each limit cycle to have a ratio ≤ 0.9. A ratio of 0 indicates that the concentration of each species reaches zero during each cell cycle.

References

[1] V. V. Isaeva. "Self-organization in biological systems". *Biology Bulletin* 39 (2012), pp. 110–118. 10.1134/S1062359012020069.

[2] B. Hess and A. Mikhailov. "Self-organization in living cells". *Science* 264 (1994), pp. 223–224.

[3] P. Richard *et al.* "Acetaldehyde Mediates the Synchronization of Sustained Glycolytic Oscillations in Populations of Yeast Cells". *European Journal of Biochemistry* 235 (1996), pp. 238–241. 10.1111/j.1432-1033.1996.00238.x.

[4] S. Danø, P. G. Sørensen, and F. Hynne. "Sustained oscillations in living cells". *Nature* 402 (1999), pp. 320–322. 10.1038/46329.

[5] J. E. Ferrell, T. Y.-C. Tsai, and Q. Yang. "Modeling the Cell Cycle: Why Do Certain Circuits Oscillate?" *Cell* 144 (2011), pp. 874–885. 10.1016/j.cell.2011.03.006.

[6] A. Murray and M. Kirschner. "Dominoes and clocks: the union of two views of the cell cycle". *Science* 246 (1989), pp. 614–621. 10.1126/science.2683077.

[7] J. J. Tyson, A. Csikasz-Nagy, and B. Novak. "The dynamics of cell cycle regulation". *BioEssays* 24 (2002), pp. 1095–1109. 10.1002/bies.10191.

[8] C. Gérard and A. Goldbeter. "Temporal self-organization of the cyclin/Cdk network driving the mammalian cell cycle". *Proceedings of the National Academy of Sciences* 106 (2009), pp. 21643–21648. 10.1073/pnas.0903827106.

[9] C. Gérard and A. Goldbeter. "From quiescence to proliferation: Cdk oscillations drive the mammalian cell cycle". *Frontiers in Physiology* 3 (2012), pp. 1–18. 10.3389/fphys.2012.00413.

[10] M. Barberis *et al.* "Sic1 plays a role in timing and oscillatory behaviour of B-type cyclins". *Biotechnology Advances* 30 (2012), pp. 108–130. 10.1016/j.biotechadv.2011.09.004.

[11] M. Barberis, E. Klipp, M. Vanoni, and L. Alberghina. "Cell Size at S Phase Initiation: An Emergent Property of the G1/S Network". *PLoS Computational Biology* 3 (2007), e64. 10.1371/journal.pcbi.0030064.

[12] R. Steuer. "Effects of stochasticity in models of the cell cycle: from quantized cycle times to noise-induced oscillations". *Journal of Theoretical Biology* 228 (2004), pp. 293–301. 10.1016/j.jtbi.2004.01.012.

[13] C. Gérard, D. Gonze, and A. Goldbeter. "Effect of positive feedback loops on the robustness of oscillations in the network of cyclin-dependent kinases driving the mammalian cell cycle". *FEBS Journal* 279 (2012), pp. 3411–3431. 10.1111/j.1742-4658.2012.08585.x.

[14] J. E. Ferrell. "Feedback loops and reciprocal regulation: recurring motifs in the systems biology of the cell cycle". *Current Opinion in Cell Biology* 25 (2013), pp. 676–686. 10.1016/j.ceb.2013.07.007.

[15] B. Ananthasubramaniam and H. Herzel. "Positive Feedback Promotes Oscillations in Negative Feedback Loops". *PLoS ONE* 9 (2014). Ed. by M. Thattai, e104761. 10.1371/journal.pone.0104761.

[16] R. Thomas. "On the Relation Between the Logical Structure of Systems and Their Ability to Generate Multiple Steady States or Sustained Oscillations". 1981, pp. 180–193. 10.1007/978-3-642-81703-8_24.

[17] N. Rangarajan *et al.* "Disorder, oscillatory dynamics and state switching: the role of c-Myc". *Journal of Theoretical Biology* 386 (2015), pp. 105–114. 10.1016/j.jtbi.2015.09.013.

[18] J. D. Moore. "In the wrong place at the wrong time: does cyclin mislocalization drive oncogenic transformation?" *Nature Reviews Cancer* 13 (2013), pp. 201–208. 10.1038/nrc3468.

[19] N. A. Cookson, S. W. Cookson, L. S. Tsimring, and J. Hasty. "Cell cycle-dependent variations in protein concentration". *Nucleic Acids Research* 38 (2010), pp. 2676–2681. 10.1093/nar/gkp1069.

[20] K. C. Chen *et al.* "Kinetic Analysis of a Molecular Model of the Budding Yeast Cell Cycle". *Molecular Biology of the Cell* 11 (2000). Ed. by M. J. Solomon, pp. 369–391. 10.1091/mbc.11.1.369.

[21] K. C. Chen *et al.* "Integrative Analysis of Cell Cycle Control in Budding Yeast". *Molecular Biology of the Cell* 15 (2004), pp. 3841–3862. 10.1091/mbc.e03-11-0794.

[22] E. Doedel, H. B. Keller, and J. P. Kernevez. "Numerical Analysis And Control of Bifurcation Problems (I): Bifurcation in Finite Dimensions". *International Journal of Bifurcation and Chaos* 01 (1991), pp. 493–520. 10.1142/S0218127491000397.

[23] V. Chickarmane, S. R. Paladugu, F. Bergmann, and H. M. Sauro. "Bifurcation discovery tool." *Bioinformatics (Oxford, England)* 21 (2005), pp. 3688–90. 10.1093/bioinformatics/bti603.

[24] J. Levering, U. Kummer, K. Becker, and S. Sahle. "Glycolytic oscillations in a model of a lactic acid bacterium metabolism". *Biophysical Chemistry* 172 (2013), pp. 53–60. 10.1016/j.bpc.2012.11.002.

[25] M. A. Savageau *et al.* "Phenotypes and tolerances in the design space of biochemical systems." *Proceedings of the National Academy of Sciences of the United States of America* 106 (2009), pp. 6435–40. 10.1073/pnas.0809869106.

[26] J. G. Lomnitz and M. A. Savageau. "Phenotypic deconstruction of gene circuitry". *Chaos: An Interdisciplinary Journal of Nonlinear Science* 23 (2013), p. 025108. 10.1063/1.4809776.

[27] J. G. Lomnitz and M. A. Savageau. "Elucidating the genotype–phenotype map by automatic enumeration and analysis of the phenotypic repertoire". *npj Systems Biology and Applications* 1 (2015), p. 15003. 10.1038/npjsba.2015.3.

[28] D. Hilbert. "Mathematical problems". *Bulletin of the American Mathematical Society* 8 (1902), pp. 437–480. 10.1090/S0002-9904-1902-00923-3.

[29] D. Battogtokh and J. J. Tyson. "Bifurcation analysis of a model of the budding yeast cell cycle". *Chaos: An Interdisciplinary Journal of Nonlinear Science* 14 (2004), pp. 653–661. 10.1063/1.1780011.

[30] C. Gérard, J. J. Tyson, D. Coudreuse, and B. Novák. "Cell Cycle Control by a Minimal Cdk Network". *PLOS Computational Biology* 11 (2015). Ed. by J. J. Saucerman, e1004056. 10.1371/journal.pcbi.1004056.

[31] J. G. Lomnitz and M. A. Savageau. "Design Space Toolbox V2: Automated Software Enabling a Novel Phenotype-Centric Modeling Strategy for Natural and Synthetic Biological Systems". *Frontiers in Genetics* 7 (2016), p. 118. 10.3389/fgene.2016.00118.

[32] C. Linke *et al.* "A Clb/Cdk1-mediated regulation of Fkh2 synchronizes CLB expression in the budding yeast cell cycle". *npj Systems Biology and Applications* 3 (2017), p. 7. 10.1038/s41540-017-0008-1.

[33] A. Pic-Taylor, Z. Darieva, B. A. Morgan, and A. D. Sharrocks. "Regulation of Cell Cycle-Specific Gene Expression through Cyclin-Dependent Kinase-Mediated Phosphorylation of the Forkhead Transcription Factor Fkh2p". *Molecular and Cellular Biology* 24 (2004), pp. 10036–10046. 10.1128/MCB.24.22.10036-10046.2004.

[34] J. Bloom and F. R. Cross. "Multiple levels of cyclin specificity in cell-cycle control". *Nature Reviews Molecular Cell Biology* 8 (2007), pp. 149–160. 10.1038/nrm2105.

[35] F. R. Cross, M. Yuste-Rojas, S. Gray, and M. D. Jacobson. "Specialization and Targeting of B-Type Cyclins". *Molecular Cell* 4 (1999), pp. 11–19. 10.1016/S1097-2765(00)80183-5.

[36] A. D. Donaldson *et al.* "CLB5-Dependent Activation of Late Replication Origins in S. cerevisiae". *Molecular Cell* 2 (1998), pp. 173–182. 10.1016/S1097-2765(00)80127-6.

[37] H. Richardson *et al.* "Cyclin-B homologs in Saccharomyces cerevisiae function in S phase and in G2". *Genes & Development* 6 (1992), pp. 2021–2034. 10.1101/gad.6.11.2021.

[38] K. Pecani and F. R. Cross. "Degradation of the mitotic cyclin clb3 is not required for mitotic exit but is necessary for G1 cyclin control of the succeeding cell cycle". *Genetics* 204 (2016), pp. 1479–1494. 10.1534/genetics.116.194837.

[39] E. Schwob and K. Nasmyth. "CLB5 and CLB6, a new pair of B cyclins involved in DNA replication in Saccharomyces cerevisiae". *Genes & Development* 7 (1993), pp. 1160–1175. 10.1101/gad.7.7a.1160.

[40] I. Fitch *et al.* "Characterization of four B-type cyclin genes of the budding yeast Saccharomyces cerevisiae". *Molecular Biology of the Cell* 3 (1992), pp. 805–818. 10.1091/mbc.3.7.805.

[41] C. Dahmann and B. Futcher. "Specialization of B-type cyclins for mitosis or meiosis in S. cerevisiae." *Genetics* 140 (1995), 957 LP –963.

[42] F. R. Cross, L. Schroeder, and J. M. Bean. "Phosphorylation of the Sic1 Inhibitor of B-Type Cyclins in Saccharomyces cerevisiae Is Not Essential but Contributes to Cell Cycle Robustness". *Genetics* 176 (2007), pp. 1541–1555. 10.1534/genetics.107.073494.

[43] R. A. Fasani and M. A. Savageau. "Automated construction and analysis of the design space for biochemical systems". *Bioinformatics* 26 (2010), pp. 2601–2609. 10.1093/bioinformatics/btq479.

[44] A. Amon, M. Tyers, B. Futcher, and K. Nasmyth. "Mechanisms that help the yeast cell cycle clock tick: G2 cyclins transcriptionally activate G2 cyclins and repress G1 cyclins". *Cell* 74 (1993), pp. 993–1007. 10.1016/0092-8674(93)90722-3.

[45] J. M. Skotheim, S. Di Talia, E. D. Siggia, and F. R. Cross. "Positive feedback of G1 cyclins ensures coherent cell cycle entry". *Nature* 454 (2008), pp. 291–296. 10.1038/nature07118.

[46] R. Verma *et al.* "Phosphorylation of Sic1p by G1 Cdk required for its degradation and entry into S phase." *Science (New York, N.Y.)* 278 (1997), pp. 455–60. 10.1126/science.278.5337.455.

[47] T. Moll *et al.* "The role of phosphorylation and the CDC28 protein kinase in cell cycle-regulated nuclear import of the S. cerevisiae transcription factor SW15". *Cell* 66 (1991), pp. 743–758. 10.1016/0092-8674(91)90118-I.

[48] S. J. Rahi *et al.* "The CDK-APC/C Oscillator Predominantly Entrains Periodic Cell-Cycle Transcription". *Cell* 165 (2016), pp. 475–487. 10.1016/j.cell.2016.02.060.

[49] M. Apri, J. Molenaar, M. de Gee, and G. van Voorn. "Efficient Estimation of the Robustness Region of Biological Models with Oscillatory Behavior". *PLoS ONE* 5 (2010). Ed. by D. Di Bernardo, e9865. 10.1371/journal.pone.0009865.

[50] F. R. Cross, V. Archambault, M. Miller, and M. Klovstad. "Testing a Mathematical Model of the Yeast Cell Cycle". *Molecular Biology of the Cell* 13 (2002). Ed. by M. J. Solomon, pp. 52–70. 10.1091/mbc.01-05-0265.

[51] S. Ghaemmaghami *et al.* "Global analysis of protein expression in yeast". *Nature* 425 (2003), pp. 737–741. 10.1038/nature02046.

[52] M. Barberis. "Sic1 as a timer of Clb cyclin waves in the yeast cell cycle - design principle of not just an inhibitor". *FEBS Journal* 279 (2012), pp. 3386–3410. 10.1111/j.1742-4658.2012.08542.x.

[53] C. B. Epstein and F. R. Cross. "CLB5: a novel B cyclin from budding yeast with a role in S phase." *Genes & Development* 6 (1992), pp. 1695–1706. 10.1101/gad.6.9.1695.

[54] A. E. Ikui and F. R. Cross. "Specific Genetic Interactions Between Spindle Assembly Checkpoint Proteins and B-Type Cyclins in Saccharomyces cerevisiae". *Genetics* 183 (2009), pp. 51–61. 10.1534/genetics.109.105148.

[55] T. Kuczera *et al.* "Dissection of mitotic functions of the yeast cyclin Clb2". *Cell Cycle* 9 (2010), pp. 2611–2619. 10.4161/cc.9.13.12082.

[56] U. Surana *et al.* "The role of CDC28 and cyclins during mitosis in the budding yeast S. cerevisiae". *Cell* 65 (1991), pp. 145–161. 10.1016/0092-8674(91)90416-V.

[57] D. A. Ball *et al.* "Stochastic exit from mitosis in budding yeast: Model predictions and experimental observations". *Cell Cycle* 10 (2011), pp. 999–1009. 10.4161/cc.10.6.14966.

[58] M. Barberis *et al.* "A low number of SIC1 mRNA molecules ensures a low noise level in cell cycle progression of budding yeast". *Molecular BioSystems* 7 (2011), pp. 2804–2812. 10.1039/c1mb05073g.

[59] D. A. Ball *et al.* "Oscillatory Dynamics of Cell Cycle Proteins in Single Yeast Cells Analyzed by Imaging Cytometry". *PLoS ONE* 6 (2011). Ed. by D. Lew, e26272. 10.1371/journal.pone.0026272.

[60] D. Ball *et al.* "Measurement and modeling of transcriptional noise in the cell cycle regulatory network". *Cell Cycle* 12 (2013), pp. 3392–3407. 10.4161/cc.26257.

[61] D. Barik, D. A. Ball, J. Peccoud, and J. J. Tyson. "A Stochastic Model of the Yeast Cell Cycle Reveals Roles for Feedback Regulation in Limiting Cellular Variability". *PLOS Computational Biology* 12 (2016). Ed. by L. You, e1005230. 10.1371/journal.pcbi.1005230.

[62] J. G. Lomnitz and M. A. Savageau. "Strategy Revealing Phenotypic Differences among Synthetic Oscillator Designs". *ACS Synthetic Biology* 3 (2014), pp. 686–701. 10.1021/sb500236e.

[63] F. S. Heldt, R. Lunstone, J. J. Tyson, and B. Novák. "Dilution and titration of cell-cycle regulators may control cell size in budding yeast". *PLOS Computational Biology* 14 (2018). Ed. by E. A. Sobie, e1006548. 10.1371/journal.pcbi.1006548.

[64] S. Rata *et al.* "Two Interlinked Bistable Switches Govern Mitotic Control in Mammalian Cells". *Current Biology* 28 (2018), 3824–3832.e6. 10.1016/j.cub.2018.09.059.

[65] M. Barberis *et al.* "CK2 regulates in vitro the activity of the yeast cyclin-dependent kinase inhibitor Sic1". *Biochemical and Biophysical Research Communications* 336 (2005), pp. 1040–1048. 10.1016/j.bbrc.2005.08.224.

[66] R. Pippa *et al.* "p27Kip1 represses transcription by direct interaction with p130/E2F4 at the promoters of target genes". *Oncogene* 31 (2012), pp. 4207–4220. 10.1038/onc.2011.582.

[67] S. Orlando *et al.* "p27 Kip1 and p21 Cip1 collaborate in the regulation of transcription by recruiting cyclin–Cdk complexes on the promoters of target genes". *Nucleic Acids Research* 43 (2015), pp. 6860–6873. 10.1093/nar/gkv593.

[68] A. Abudukelimu, T. D. Mondeel, M. Barberis, and H. V. Westerhoff. "Learning to read and write in evolution: from static pseudoenzymes and pseudosignalers to dynamic gear shifters". *Biochemical Society Transactions* 45 (2017), pp. 635–652. 10.1042/BST20160281.

[69] T. D. Mondeel *et al.* "Maps for when the living gets tough: Maneuvering through a hostile energy landscape". *IFAC-PapersOnLine* 49 (2016), pp. 364–370. 10.1016/j.ifacol.2017.03.002.

[70] R. A. Weinberg. *The biology of cancer*. English. 2nd. Garland Science, Taylor & Francis Group, New York and London, 2014.

[71] T. Kluyver *et al.* "Jupyter Notebooks—a publishing format for reproducible computational workflows". *Positioning and Power in Academic Publishing: Players, Agents and Agendas* (2016), pp. 87–90. 10.3233/978-1-61499-649-1-87.

[72] R. L. Rossi *et al.* "Subcellular Localization of the Cyclin Dependent Kinase Inhibitor Sic1 is Modulated by the Carbon Source in Budding Yeast". *Cell Cycle* 4 (2005), pp. 1798–1807. 10.4161/cc.4.12.2189.

[73] S. Hoops *et al.* "COPASI–a COmplex PAthway SImulator". *Bioinformatics* 22 (2006), pp. 3067–3074. 10.1093/bioinformatics/btl485.

[74] A. Dhooge, W. Govaerts, and Y. a. Kuznetsov. "MATCONT". *ACM Transactions on Mathematical Software* 29 (2003), pp. 141–164. 10.1145/779359.779362.

[75] A. Dhooge *et al.* "New features of the software MatCont for bifurcation analysis of dynamical systems". *Mathematical and Computer Modelling of Dynamical Systems* 14 (2008), pp. 147–175. 10.1080/13873950701742754.

[76] B. Ermentrout. *Simulating, analyzing, and animating dynamical systems: a guide to XPPAUT for researchers and students*. Vol. 14. Siam, 2002.

[77] B. N. Kholodenko, O. V. Demin, and H. V. Westerhoff. "Control Analysis of Periodic Phenomena in Biological Systems". *The Journal of Physical Chemistry B* 101 (1997), pp. 2070–2081. 10.1021/jp962336u.

[78] K. A. Reijenga, H. V. Westerhoff, B. N. Kholodenko, and J. L. Snoep. "Control Analysis for Autonomously Oscillating Biochemical Networks". *Biophysical Journal* 82 (2002), pp. 99–108. 10.1016/S0006-3495(02)75377-0.

[79] D. A. Fell. "Metabolic control analysis: a survey of its theoretical and experimental development." *The Biochemical journal* 286 (Pt 2 (1992), pp. 313–30.

[80] H. V. Westerhoff. "Signalling control strength". *Journal of Theoretical Biology* 252 (2008), pp. 555–567. 10.1016/j.jtbi.2007.11.035.

[81] M. Domijan, P. E. Brown, B. V. Shulgin, and D. A. Rand. "PeTTSy: a computational tool for perturbation analysis of complex systems biology models". *BMC Bioinformatics* 17 (2016), p. 124. 10.1186/s12859-016-0972-2.

[82] R. Wäsch and F. R. Cross. "APC-dependent proteolysis of the mitotic cyclin Clb2 is essential for mitotic exit". *Nature* 418 (2002), pp. 556–562. 10.1038/nature00856.

[83] B. Teusink, B. M. Bakker, and H. V. Westerhoff. "Control of frequency and amplitudes is shared by all enzymes in three models for yeast glycolytic oscillations". *Biochimica et Biophysica Acta (BBA) - Bioenergetics* 1275 (1996), pp. 204–212. 10.1016/0005-2728(96)00026-6.

[84] D. Fiedler *et al.* "Functional Organization of the S. cerevisiae Phosphorylation Network". *Cell* 136 (2009), pp. 952–963. 10.1016/j.cell.2008.12.039.

[85] R. Visintin *et al.* "The phosphatase Cdc14 triggers mitotic exit by reversal of Cdk-dependent phosphorylation." *Molecular cell* 2 (1998), pp. 709–18.

[86] M. A. Savageau and E. O. Voit. "Recasting nonlinear differential equations as S-systems: a canonical nonlinear form". *Mathematical Biosciences* 87 (1987), pp. 83–115. 10.1016/0025-5564(87)90035-6.

[87] E. O. Voit. "Biochemical Systems Theory: A Review". *ISRN Biomathematics* 2013 (2013), pp. 1–53. 10.1155/2013/897658.

[88] M. A. Savageau. "Introduction to S-systems and the underlying power-law formalism". *Mathematical and Computer Modelling* 11 (1988), pp. 546–551. 10.1016/0895-7177(88)90553-5.

[89] P. C. Hollenhorst *et al.* "Forkhead Genes in Transcriptional Silencing, Cell Morphology and the Cell Cycle: Overlapping and Distinct Functions for FKH1 and FKH2 in Saccharomyces cerevisiae". *Genetics* 154 (2000), pp. 1533–1548. 10.1093/genetics/154.4.1533.

[90] J. A. Ubersax *et al.* "Targets of the cyclin-dependent kinase Cdk1". *Nature* 425 (2003), pp. 859–864. 10.1038/nature02062.

[91] F. M. Yeong, H. H. Lim, Y. Wang, and U. Surana. "Early Expressed Clb Proteins Allow Accumulation of Mitotic Cyclin by Inactivating Proteolytic Machinery during S Phase". *Molecular and Cellular Biology* 21 (2001), pp. 5071–5081. 10.1128/MCB.21.15.5071-5081.2001.

[92] R. Kumar *et al.* "Forkhead transcription factors, Fkh1p and Fkh2p, collaborate with Mcm1p to control transcription required for M-phase". *Current Biology* 10 (2000), pp. 896–906. 10.1016/S0960-9822(00)00618-7.

[93] D. Reynolds *et al.* "Recruitment of Thr 319-phosphorylated Ndd1p to the FHA domain of Fkh2p requires Clb kinase activity: a mechanism for CLB cluster gene activation." *Genes & development* 17 (2003), pp. 1789–802. 10.1101/gad.1074103.

[94] A. D. Rudner and A. W. Murray. "Phosphorylation by Cdc28 Activates the Cdc20-Dependent Activity of the Anaphase-Promoting Complex". *The Journal of Cell Biology* 149 (2000), pp. 1377–1390. 10.1083/jcb.149.7.1377.

[95] M. Shirayama, A. Tóth, M. Gálová, and K. Nasmyth. "APC(Cdc20) promotes exit from mitosis by destroying the anaphase inhibitor Pds1 and cyclin Clb5". *Nature* 402 (1999), pp. 203–207. 10.1038/46080.

[96] W. Zachariae. "Control of Cyclin Ubiquitination by CDK-Regulated Binding of Hct1 to the Anaphase Promoting Complex". *Science* 282 (1998), pp. 1721–1724. 10.1126/science.282.5394.1721.

[97] P. Nash *et al.* "Multisite phosphorylation of a CDK inhibitor sets a threshold for the onset of DNA replication". *Nature* 414 (2001), pp. 514–521. 10.1038/35107009.

CHAPTER 3

ChIP-exo analysis highlights Fkh1 and Fkh2 transcription factors as hubs that integrate multi-scale networks in budding yeast

3.1 Introduction 86
3.2 Materials and Methods 88
Yeast strains and growth conditions 88
ChIP-exo 88
Data processing 89
Gene annotation and data analysis 90
KEGG pathway map visualization 91
3.3 Results 91
Data analysis pipeline using the novel *maxPeak* method 91
Consensus of verified and novel targets of Fkh1 and Fkh2 . . . 92
The correlation between Fkh and target expression levels 96
Dynamics of cell cycle-regulated target genes 97
Functional enrichment of identified Fkh target genes 99
Fkh targets in their functional context 101
3.4 Discussion 103
Supplementary Materials and Methods 111
Supplementary Text 115

Adapted from:
T.D.G.A. Mondeel, P. Holland, J. Nielsen, M. Barberis, ChIP-exo analysis highlights Fkh1 and Fkh2 transcription factors as hubs that integrate multi-scale networks in budding yeast, Nucleic Acids Res. 47 (2019) 7825–7841. 10.1093/nar/gkz603

Closely related work:
P. Holland, J. Nielsen, T.D.G.A. Mondeel, M. Barberis, Coupling Cell Division to Metabolic Pathways Through Transcription, in: Encycl. Bioinforma. Comput. Biol., Elsevier, 2019: pp. 74–93. 10.1016/B978-0-12-809633-8.20081-2

M. Barberis, T.D.G.A. Mondeel, Unveiling Forkhead-mediated regulation of yeast cell cycle and metabolic networks, Comput. Struct. Biotechnol. J. (2022). In Press.

> "A key aim of postgenomic biomedical research is to systematically catalogue all molecules and their interactions within a living cell. There is a clear need to understand how these molecules and the interactions between them determine the function of this enormously complex machinery, both in isolation and when surrounded by other cells."
>
> — Albert-László Barabási & Zoltán N. Oltvai [1]

Abstract

The understanding of the multi-scale nature of molecular networks represents a major challenge. For example, regulation of a timely cell cycle must be coordinated with growth, during which changes in metabolism occur, and integrate information from the extracellular environment, e.g. signal transduction. Forkhead transcription factors are evolutionarily conserved among eukaryotes, and may coordinate a timely cell cycle progression in budding yeast. Specifically, Fkh1 and Fkh2 are expressed during a lengthy window of the cell cycle, thus are potentially able to function as hubs in the multi-scale cellular environment that interlock various biochemical networks. Here we report on a novel ChIP-exo dataset for Fkh1 and Fkh2 in both exponential and stationary phases, which is analyzed using novel and existing software tools. Our analysis confirms known Forkhead targets from available ChIP-chip studies and highlights novel ones involved in the cell cycle, metabolism and signal transduction. Target genes are analyzed with respect to their function, temporal expression during the cell cycle, correlation with Fkh1 and Fkh2 as well as signaling and metabolic pathways they occur in. Furthermore, differences in targets between Fkh1 and Fkh2 are presented. Our work highlights Forkhead transcription factors as hubs that integrate multi-scale networks to achieve proper timing of cell division in budding yeast.

3.1 Introduction

Biological systems exploit their functions across space and time, and their robustness results from the coherent integration of functionally diverse elements (e.g., molecules, modules) that interact selectively and nonlinearly [2]. Thus, the cross-talk between modules representing cellular layers of regulation (e.g., gene regulation, cell cycle, metabolism, signal transduction) is crucial to achieve system's functions. In this context, identification of elements with high connectivity (hubs) bridging multiple spatial, temporal and functional scales within cellular networks is an important challenge in Systems Biology. This also holds for the generation of multi-scale models with the aim of understanding how a function emerges from a network of interactions [3].

Transcription factors are pivotal in gene regulation, by switching on or off entire molecular pathways, thus modulating their activity or, more subtly, affecting the timing and extent of their activation. Among these regulators, Forkhead (Fkh)

transcription factors (Forkhead Box (FOX) in mammals) are highly conserved across eukaryotes, and have garnered interest because of their involvement in multiple cellular pathways that, when dysregulated, may lead to development of pathologies such as neurodegeneration, cancer, and aging [4–7].

The homologues of the FOX proteins in budding yeast, Fkh1 and Fkh2, play an important role as regulators of the *CLB2*-cluster, i.e. a set of genes transcriptionally regulated after CLB2 activation [8]. This set consists of 33 genes whose transcription peaks in late G2/early M phase of the cell cycle [8]. Fkh2 promotes activation of the *CLB2* promoter, in complex with the Mcm1 scaffold protein and the co-activator Ndd1, leading to cell division [9–12]. Fkh1 function overlaps with that of Fkh2, but it binds less efficiently to the *CLB2* promoter and represses *CLB2* transcription [13–15].

We have recently demonstrated that Fkh2 synchronizes the temporal expression of mitotic *CLB* genes by connecting the cyclins *CLB5*, *CLB3* and *CLB2* in a linear cascade, and ensuring their timely activation [16]. We also showed, as have other [17], an Ndd1/Fkh interaction, but the function of the Ndd1/Fkh complex is currently not understood. Fkh1 is expressed during S and G2 phases, and its transcript levels peak in the S phase, whereas Fkh2 is expressed from G1 until the M phase, and its transcript levels peak during the G1(P) (pre-replicative G1) and S phase [18, 19]. This relatively lengthy window of expression, in particular for Fkh2, may allow the Fkh's to interact with a diverse set of temporally separated cellular pathways.

We have also found a possible pathway for an interplay between metabolism and cell cycle, with the NAD^+-dependent histone deacetylases Sir2 modulating the Fkh-dependent regulation of target genes [20]. Sir2 associates with Fkh in the G1 and M phases, where it inhibits activation of *CLB2* through Fkh-mediated binding to the *CLB2* promoter [20]. The NAD^+/NADH ratio reflects the intracellular redox state, and is a readout of metabolic activity [21].

Additional data also suggest a possible role of Fkh in cellular processes beside cell cycle regulation. Microarray-based RNA profiling identified four target genes of Fkh1 and two targets of Fkh2 [22]. Furthermore, chromatin immunoprecipitation (ChIP)-based methodologies, specifically ChIP-chip [23], have retrieved hundreds of targets of Fkh1 and Fkh2 [24–26]. Moreover, ChIP-chip-based computational strategies to identify sequence patterns that bind to transcription factors (referred to as binding motifs) have identified similar binding motifs for Fkh1 and Fkh2 [26], as also reported in the YeTFaSCo database [27].

These studies identified several potential Fkh targets in metabolism. For example, Fkh1 has been suggested to regulate *FAB1*, which encodes a vacuolar membrane kinase that generates phosphatidylinositol – the latter involved in vacuolar sorting and homeostasis – and *ALG5*, which encodes a beta-glucosyltransferase that is involved in asparagine-linked glycosylation in the endoplasmic reticulum [24–26]. Similarly, Fkh2 has been suggested to regulate several metabolic enzymes such as *GLN1*, encoding a glutamine synthetase; *IDI1*, encoding an isopentenyl diphosphate that catalyzes an essential activation step in the isoprenoid biosynthetic pathway; and *UTH1*, encoding a mitochondrial inner membrane protein implicated in cell wall biogenesis [24–26]. Furthermore,

HOS3, encoding a histone deacetylase, has been shown as a common enzymatic target of Fkh1 and Fkh2 [24–26]. Together, this evidence suggests a Fkh-mediated connectivity between cell cycle and metabolism.

Here, we provide a comprehensive, up-to-date overview of the current knowledge of Fkh target genes. First, we report on a novel dataset of Fkh targets using ChIP-exo, which combines ChIP with lambda exonuclease digestion followed by high-throughput sequencing, that allows identification of a nearly complete set of binding sites at near single nucleotide resolution [28]. We have recently employed ChIP-exo to investigate targets of transcription factors in budding yeast [29, 30]. The ChIP-exo dataset generated in this study was annotated using GEMMER, a web-based data-integration and visualization tool that we have recently developed to integrate and visualize the large experimental data available for budding yeast [31]. Subsequently, known and novel Fkh target genes were analyzed with respect to their function, temporal expression during the cell cycle as well as signaling and metabolic pathways they occur in. Emphasis is given to targets connecting cell cycle with other cellular processes, in particular metabolism. Our study clarifies and expands the understanding on the role that Fkh have as hubs that integrate multi-scale regulatory networks to achieve proper timing of cell division.

3.2 Materials and Methods

Yeast strains and growth conditions

The yeast strain BY4741 (*MATa his3Δ1 leu2Δ0 met15Δ0 ura3Δ0*) was used to generate the respective strains Fkh1-Myc (*FKH1-MYC9::kanMX6*) and Fkh2-Myc (*FKH2-MYC9::kanMX6*), as described [20]. Yeast strains were grown on plates with YPD with G418 (Formedium) 200 mg/L or in liquid cultures of defined media containing NH_4SO_4 3.75 g/L, KH_2PO_4 7.18 g/L, $MgSO_4$ 0.25 g/L, Glucose 10 g/L, Complete supplement mix (Formedium, DCS0019 - Adenine 5 mg/L, L-Arg 25 mg/L, L-Asp 40 mg/L, L-His 10 mg/L, L-Iso 25 mg/L, L-Leu 50 mg/L, L-Lys 25 mg/L, L-Met 10 mg/L, l-Phe 25 mg/L, L-Thr 50 mg/L, L-Trp 25 mg/L, L-Tyr 25 mg/L, Uracil 10 mg/L, Val 70 mg/L), Vitamin solution (D-Biotin 0.05 mg/L, D-Pantothenic acid 1 mg/L, Thiamin-HCl 1 mg/L, Pyridoxin-HCl 1 mg/L, Nicotinic acid 1 mg/L, 4-Aminobenzoic acid 0.2 mg/L, myo-inositol 25 mg/L) and Trace metal solution ($FeSO_4$ 3 mg/L, $ZnSO_4$ 4.5 mg/L, $CaCl_2$ 4.5 mg/L, $MnCl_2$ 0.84 mg/L, $CoCl_2$ 0.3 mg/L, $CuSO_4$ 0.3 mg/L, Na_2MoO_4 0.4 mg/L, H_3BO_3 1.0 mg/L, KI 0.1 mg/L, Na_2EDTA 19 mg/L). pH of defined media was adjusted to 6.35 by adding KOH.

ChIP-exo

To start the liquid cultures, a yeast colony carrying Fkh1-Myc or Fkh2-Myc was picked mid-day into the above defined media and cultured with shaking at $30\,^{\circ}\mathrm{C}$ until the next morning. Cultures were then split to become exponential and stationary phase cultures. Cultures in exponential phase were started at $OD_{600} \sim 0.2$

and grown until OD600 for Fkh1 replicates: 0.75, 0.72 and Fkh2 replicates: 0.80, 0.80. Cultures in stationary phase were grown until the afternoon and collected until OD_{600} for Fkh1 replicates: 2.00, 1.70 and Fkh2 replicates: 1.76, 1.78. For the ChIP-exo experiments, 100 OD units of cells were collected from each culture, diluted to $OD_{600} \sim 0.7$ with water, supplemented with formaldehyde (Sigma F8775) to a final concentration of 1% and left shaking at room temperature for 15 min. Glycine (Sigma G7126) was added to quench the cross linking at a final concentration of 125 mM and left shaking for 5 min. Cells were then washed twice with cold TBS (Tris-HCl (Sigma 252859) pH 7.5 1 mM, NaCl (Sigma, S3014) 150 mM) and snap frozen in liquid N_2. ChIP-exo was performed according to the original protocol [28] with modifications as described [32].

Data processing

Raw reads were mapped to the SacCer3 genome[1], downloaded from the Saccharomyces Genome Database (SGD) website[2] with Bowtie2 [33]. SAM files were converted to BAM files, sorted and indexed using SAMtools 1.3.1 [34]. ChIP-exo data analysis was performed through a pipeline that uses two existing software tools and a novel method, which we refer to as *maxPeak*, for peak detection. The existing tools, GEM (*G*enome wide *E*vent finding and *M*otif discovery) [35] and MACE (*M*odel based *A*nalysis of *C*hIP-*E*xo) [36], require the sorted and indexed BAM files as input and use iteration schemes to identify and enrich peaks. Data analysis by GEM and MACE was performed through the command line. GEM and MACE require a relatively large amount of strong peaks to iterate successfully. MACE was able to analyze the Fkh1 data but unable to iterate on the Fkh2 data due to the relatively low number of peaks detected; it detected only 25 strong peaks (called 'elite border pairs') for the Fkh2 data, while it requires more than 30 by default. Therefore, in order to analyze the Fkh2 data, the threshold was reduced to 25 elite border pairs. This choice comes at the cost of a higher potential for picking up noise and low quality binding events.

The *maxPeak* peak detection method was applied, starting from the indexed BAM files, using a combination of bash scripts and R scripts. Based on the principle of ChIP-exo, there is a transcription factor-specific optimal read length, where the whole binding site is covered by reads on both DNA strands, that corresponds to the width of the DNA covered by the transcription factor. We identified this read length by comparing the raw read alignments for several genes exhibiting a strong peak. We observed that a read length of 12 bp corresponds well to the strong peaks for both Fkh1 and Fkh2 (data not shown). This is consistent with the previously identified binding motifs for Fkh1 and Fkh2, which have been reported to cover a length of 8 bp and 7 bp, respectively [26], as the ChIP-exo read length is slightly larger than the binding motif due to additional "head room" that the exonuclease cannot reach.

In the *maxPeak* method, the number of reads on both + and – DNA strands

[1] S288C_reference_sequence_R64-2-1_20150113.fsa

[2] See: `https://downloads.yeastgenome.org/sequence/S288C_reference/genome_releases/`

was summed up genome wide for each nucleotide position. At this stage, biological duplicates were averaged. Finally, by using the R environment for statistical computing and graphics, the 65^{th} percentile of the maximum read counts for genes that had a maximum > 0 was calculated for each experimental condition (exponential and stationary phases) independently, creating one noise threshold level per experiment. The highest read count per gene was then divided by the noise threshold for each experiment to calculate the signal-to-noise ratio (SNR). Essentially, maxPeak ranks genes based on their signal intensity. The 65^{th} percentile normalization threshold is irrelevant for the ranking of the genes, and it only serves to set a rough threshold below which a gene's signal is considered as noise. We did not average the read counts among the experiments of each Fkh transcription factor because a significantly higher signal in the stationary phase experiments for both Fkh1 and Fkh2 was observed. This evidence suggests that there was no equal background noise across the different conditions, and that averaging may result in retrieving false positives as a consequence of the lowered threshold for the stationary phase experiments..

To score the significance of the target genes retrieved, *maxPeak* and GEM assign SNRs, whereas MACE assigns p-values. A comparison between the principles behind *maxPeak,* GEM and MACE methods is in Supplementary Materials and Methods, Fig. S3.1, whereas the specific thresholds used for each peak detection method (PDM) are indicated in Fig. 3.1 and in Supplementary Materials and Methods, Fig. S3.2–S3.4. The scripts used for data processing and the unannotated output from *maxPeak*, GEM and MACE are available as Supplementary Code Repository.

Gene annotation and data analysis

In budding yeast, the median promoter length is 455 bp [37]. To also cover the promoter regions that are longer than this median length, we have recently considered a window length of 1,000 bp [29, 30, 32]. In this work, we analyzed the data for binding enrichment up to 1,000 bp upstream of the start of 7,217 ORFs (Open Reading Frames) annotated in the sacCer3 genome, possibly reaching the coding sequence of an upstream gene. Gene annotation was performed through GEMMER, a novel web-based data-integration and visualization tool that we have recently developed for budding yeast [31] (Supplementary Materials and Methods). We retrieved annotation from GEMMER for the ~ 6800 protein-coding genes, as identified by SGD. As in GEMMER, we considered genes that have an annotated E.C. number to be enzymes; occasionally, we referred to enzymes that catalyze reactions in the Yeast 7.6 metabolic map [38, 39] as metabolic enzymes, to emphasize their specific function. The SNRs and p-values assigned by *maxPeak*, GEM and MACE were all merged into one dataset together with the annotation (Supplementary Excel Table S1). Data analysis was performed on the processed and annotated dataset described above using Python 3.6 and the Pandas and Matplotlib modules. A collection of Python scripts reproducing the data integration and Jupyter notebooks reproducing the data analysis are available in the Supplementary Code Repository and as part of a Github repository

(https://github.com/barberislab/ChIP-exo_Fkh1_Fkh2).

KEGG pathway map visualization

We used the R library Pathview to superimpose the experimental data on KEGG pathway maps [40]. We performed the mapping two times, first by mapping the set of targets identified in our experiments. Second, we associated each gene with a verification score and an associated color: (i) a value of -1 (yellow) for genes not suggested as a target by our ChIP-exo experiments, but shown by one or more of the available ChIP-chip studies [24–26]; (ii) a value of 0 (red) for target genes identified only by one of our ChIP-exo experiments, or (iii) a value of +1 (green) for target genes identified by at least one ChIP-chip study and our ChIP-exo experiments. The R script to reproduce the image generation (see Supplementary Code Repository, and images (Supplementary KEGG Figures) are available.

3.3 Results

Data analysis pipeline using the novel *maxPeak* method to detect high-confidence targets

ChIP-exo experiments were performed on Fkh1 and Fkh2, in exponential and stationary phases, for a total of four experiments (see Materials and Methods). Subsequently, two existing peak detection methods (PDMs) were applied to the ChIP-exo datasets: GEM [35] and MACE [36]. We observed a significant divergence in the target genes retrieved when comparing GEM SNR ≥ 1 and MACE (p-value ≤ 0.01) (Supplementary Text, Fig. S3.5). The large number of targets retrieved only by GEM or only by MACE led us to develop a novel ChIP-exo data analysis method, which we have named *maxPeak*, which does not use iteration and is not sensitive to a relatively low number of strong peaks. Application of three PDMs simultaneously on the ChIP-exo dataset allowed us to identify genes that are consistently retrieved as targets by Fkh1 and Fkh2 across multiple PDMs. In order for a target gene to be retrieved, it had to score above (GEM and *maxPeak*) or below (MACE) threshold in *at least* two out of three PDMs. To set thresholds that define which genes are considered targets by each of the three PDMs, we generated three 2x2 score comparisons (see Fig. 3.1 and Supplementary Materials and Methods, Fig. S3.2–S3.4). We considered any target gene that is retrieved as significant by both GEM and MACE as a confident target. Consequently, we set the threshold of significance for *maxPeak* to the lowest score obtained across all four experimental conditions by any gene that was retrieved by both GEM and MACE. Altering the 65^{th} percentile normalization threshold that we applied for maxPeak (see Materials and Methods) would not have an impact on the set of the retrieved target genes. By following this approach, we could then use *maxPeak* to discriminate between those target genes that are retrieved by only GEM or only MACE. The overlap of target genes among the three PDMs is shown in Supplementary Text, Fig. S3.6. Fig. 3.2 shows the data processing pipeline implementing the three PDMs: GEM, MACE and the novel *maxPeak*.

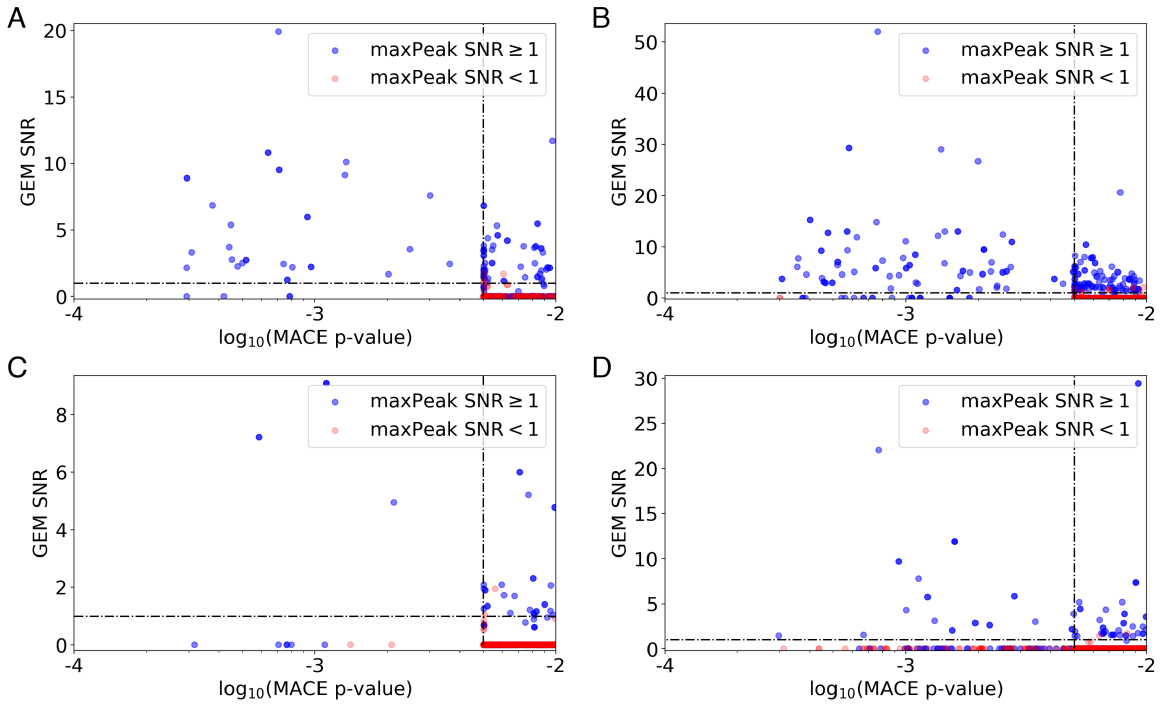


Figure 3.1: Comparison of GEM SNRs vs. MACE SNRs for target genes that scored a MACE p-value lower than 0.01. (**A**) Fkh1, exponential phase. (**B**) Fkh1, stationary phase. (**C**) Fkh2, exponential phase. (**D**) Fkh2, stationary phase. The horizontal and vertical black dotted lines represent the GEM and MACE target thresholds, respectively. Blue circles represent genes that were assigned a SNR ≥ 1 by *maxPeak*. Red circles represent genes that were assigned a SNR < 1. Genes with circles in the upper-left quadrant of each panel were considered targets. Blue circles in the bottom-left and upper-right quadrants of each panel were also considered targets.

ChIP-exo identifies a consensus of verified and novel targets of Fkh1 and Fkh2

The pipeline presented in Fig. 3.2 identified several hundred target genes of Fkh1 and Fkh2. An overview of the number of target genes that were retrieved in the four ChIP-exo experiments is reported in Table 3.1 and the targets are listed per experimental condition in Supplementary Excel Table S2. A higher number of Fkh1 targets was retrieved as compared to Fkh2 targets, and a higher number of Fkh targets was retrieved in stationary phase as compared to exponential phase. *CLB2* is considered to be the major Fkh target gene; thus, it has been considered as a positive control for both Fkh1 and Fkh2. *CLB2* was not considered significant as a Fkh1 target in exponential phase by both GEM and MACE; hence this gene was not considered a target for subsequent analyses. Conversely, in the other three experimental conditions, *CLB2* was retrieved as a Fkh target. Specifically, in all ChIP-exo experiments, *CLB2* revealed a SNR > 2 (Supplementary Excel Table S2) assigned by the *maxPeak* method. Notably, in Fkh2 datasets, *CLB2* scores the 4$^{\text{th}}$ highest SNR in exponential phase and the highest SNR in stationary phase. These

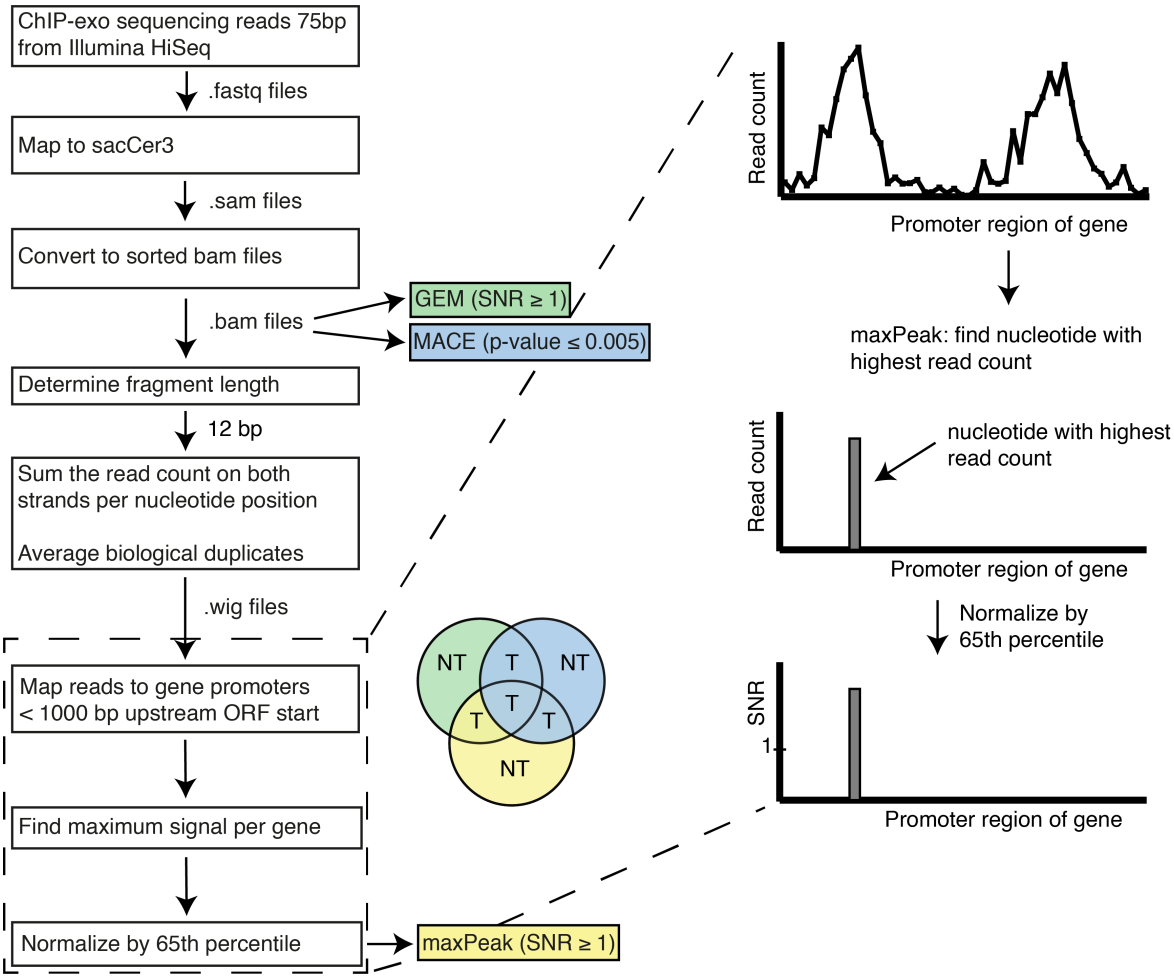


Figure 3.2: Illustration of the pipeline implemented for the identification of target genes from ChIP-exo data. First, BAM files were generated, sorted and indexed, on which GEM and MACE are run. For the *maxPeak* peak detection method, the number of reads on both DNA strands for each nucleotide is counted and, subsequently, the highest read count at a single nucleotide per gene is assigned as the gene's signal. Finally, the read count for each gene is normalized by the 65^{th} percentile of all genes with a read count > 0, calculating a signal-to-noise ratio (SNR). Finally, target genes (indicated by a 'T' in the pie chart) are selected if these are retrieved as significant by at least two out of three peak detection methods (PDMs); conversely, target genes that are retrieved as significant by only one PDM (indicated as 'NT' in the pie chart) are not considered further in the analyses.

results agree with *CLB2* being the pivotal Fkh2 target required for cell division [9–12].

A subset of target genes scored above threshold in all three PDMs (referred to as '3x PDM verified' in Table 3.1); the detailed list of targets for each experiment is reported in Supplementary Excel Table S3 and Supplementary Text, Table S1. A number of available genome-wide studies provide datasets of Fkh target genes [22, 24–26]. We focused specifically on the previous ChIP-chip studies [24–

Target genes	Fkh1 exponential	Fkh1 stationary	Fkh2 exponential	Fkh2 stationary
Total	291	416	105	220
4x ChIP verified	29	-	15	-
3x PDM verified	31	84	6	25
Novel	43	-	38	-
Cell cycle-regulated	84	122	46	65
Enzymes	60 (31)	103 (51)	18 (10)	50 (27)

Table 3.1: Number of target genes identified in this study for specific subgroups. '4x ChIP verified', verified targets retrieved by our ChIP-exo experiments and three available ChIP-chip studies. '3x PDM verified', targets retrieved by three peak detection methods, PDMs (*maxPeak*, GEM and MACE). 'Novel', novel targets retrieved by this study but not by the three available ChIP-chip studies. 'Cell-cycle regulated', targets retrieved by this study that have been described as cell cycle-regulated genes [18]. 'Enzymes', targets retrieved by this study that are enzymes; specifically, the number of enzymes that catalyze reactions in the Yeast 7.6 metabolic map (i.e. metabolic enzymes) are indicated within parentheses. Since the available ChIP-chip studies were performed in exponential phase, verified and novel targets are not available for the stationary phase experiments.

26], where experiments were performed after growing cells in exponential phase: MacIsaac et al. to an OD ~ 0.8 [26] (the experimental work was originally performed in [41], Venters et al. to an OD ~ 1.0 [25], and Ostrow et al. to an OD ~ 0.8 [24]. For this reason, for the comparison of our ChIP-exo datasets with the ChIP-chip studies, the experiments performed in stationary phase were neglected.

We quantified the agreement between the ChIP-exo peak locations and the enriched regions identified in the most recent ChIP-chip dataset [24] by overlaying them on chromosome-wide summary plots and counting the overlap (see Supplementary Text). In Fig. 3.3, a summary plot for Fkh2 in exponential phase is shown for chromosome XVI, which contains the *CLB2* gene, a major Fkh2 target (see Supplementary exo-chip Figures for all summary plots). We observed that, in exponential phase, 81% and 59% of the ChIP-exo target genes show peaks within enriched windows identified by the ChIP-chip experiments for Fkh1 and Fkh2, respectively (see Supplementary Text, Table S3.2). The remaining 19% and 41% of the ChIP-exo target genes are peak locations upstream of ORFs that the ChIP-chip study did not identify. Vice versa, 51% and 46% of the enriched ChIP-chip regions upstream of ORFs for Fkh1 and Fkh2, respectively, contain at least one significant peak event (in any PDM) as identified by ChIP-exo using our PDM thresholds (see Supplementary Text, Table S3.3). These results highlight the increased specificity achieved using ChIP-exo as compared to ChIP-chip, and the higher stringency applied by (i) the thresholds used in this work and (ii) the requirement of passing the threshold in at least 2 PDMs.

To highlight new targets of Fkh1 and Fkh2 identified using ChIP-exo, we compared the overlap between our ChIP-exo targets and the ChIP-chip targets

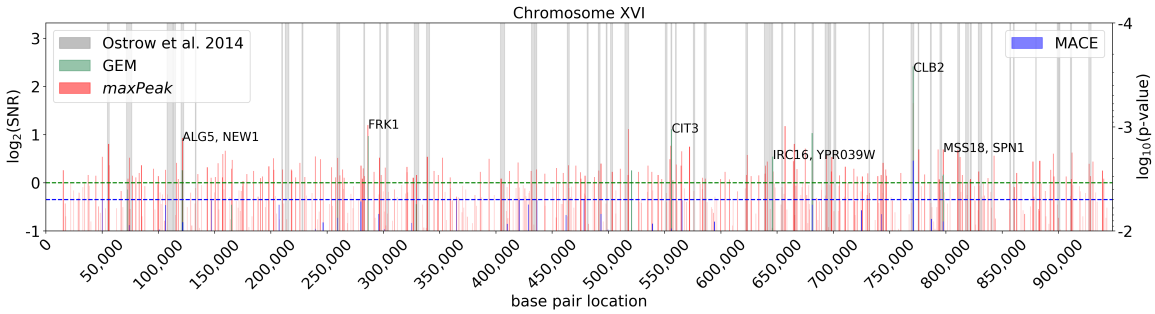


Figure 3.3: Comparison of ChIP-exo peak locations as identified by three different PDMs (maxPeak, GEM and MACE) and the ChIP-chip enriched regions identified by Ostrow et al. for Fkh2 in exponential phase on chromosome XVI. The horizontal green dotted line indicates the threshold for GEM and maxPeak; the horizontal blue dotted line indicates the threshold for MACE. All ChIP-exo peak locations with a SNR $> \frac{1}{2}$ (for GEM and maxPeak) and/or a p-value < 0.01 (MACE) are displayed. ChIP-exo target gene peaks are labeled as identified through the pipeline reported in Fig. 3.2. When multiple gene names are comma-separated in one label, the peak location was within a window of 1,000 bp upstream of all listed gene ORFs.

from [24–26]. Strikingly, only 42 out of 2939 Fkh1 target genes and 18 out of 1553 Fkh2 target genes are in common between the three published ChIP-chip studies (see Supplementary Text, Fig. S3.7 and Supplementary Excel Table S4). This lack of overlap among ChIP-chip studies is a general observation; for this reason, the recently developed ChIP-exo methodology may help to clarify these discrepancies. Indeed, our ChIP-exo experiments recovered the majority of the target genes retrieved by all three ChIP-chip studies. Furthermore, it highlights a number of novel, previously not detected, Fkh target genes. Table 3.1 summarizes the number of verified and novel target genes. The verified, thus highly reproducible, target genes by all four ChIP experiments are 29 for Fkh1 (*ADD37*, *ALG5*, *ATG42*, *BDF1*, *BUD4*, *CDS1*, *CIK1*, *DIN7*, *DSE1*, *DYN1*, *EGO2*, *ERS1*, *ESP1*, *FHL1*, *HOS3*, *JSN1*, *KIP2*, *MKK2*, *NEW1*, *RHO4*, *RPN11*, *SPC24*, *SSO2*, *SUB2*, *TDA7*, *TEL2*, *VTI1*, *YBR138C*, *YPI1*) and 15 for Fkh2 (*ATG42*, *BUD4*, *CDC20*, *CHS2*, *IRC8*, *JSN1*, *MTC6*, *PPN1*, *SCO1*, *SPO12*, *SUR7*, *SWI5*, *UTH1*, *YHP1*, *YML053C*) (see Fig. 3.4 and Supplementary Excel Table S5). Among these common target genes, 8 (for Fkh1) and 4 (for Fkh2) are enzymes. Three target genes are in common among both Fkh: *ATG42*, coding a vacuolar carboxypeptidase; *BUD4*, coding for a protein that has a role in bud site selection and is a substrate of the Clb2/Cdk1 kinase; and *JSN1*, coding an RNA-binding protein that interacts with mRNAs of membrane-associated proteins of the mitochondria. Strikingly, a potential metabolic role of Fkh target genes is suggested by the Fkh1 targets *CDS1*, coding a phosphatidate cytidylyltransferase involved in the synthesis of all major yeast phospholipids, and *ERS1*, coding a cysteine transport protein that localizes to membranes of organelles, and by the Fkh2 target the *CHS2*, encoding a chitin synthase required for chitin synthesis prior to cell division. Moreover, a

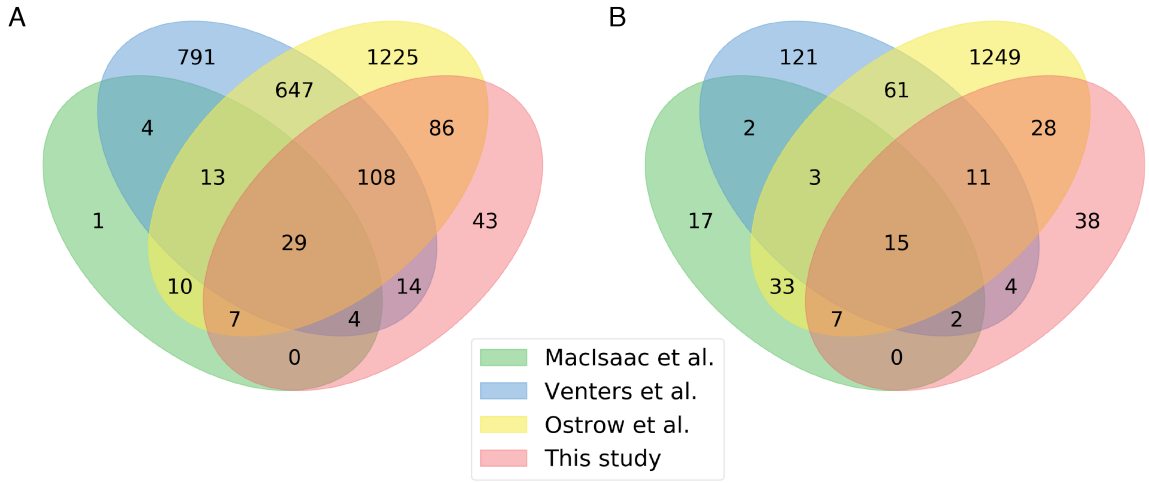


Figure 3.4: Reproduction of novel and verified Fkh target genes. 4-way Venn diagrams for Fkh1 (**A**) and Fkh2 (**B**) showing the overlap between ChIP-exo datasets and previous ChIP-chip studies that have identified Fkh target genes [24–26].

subset of target genes highlight the known role that Fkh2 plays in the control of cell division: *SWI5*, coding for the transcription factor of *SIC1* - SIC1 is the stoichiometric inhibitor of mitotic cyclin/Cdk1 kinase activities -; *CDC20*, activator of the anaphase-promoting complex/cyclosome (APC/C) required for the metaphase/anaphase transition; and *BUD4* (described earlier). Furthermore, the Fkh1 target *MKK2*, coding for a MAP kinase kinase (MAPKK) involved in the protein kinase C signaling pathway and in the control of cell integrity, points to a potential role in signal transduction. Finally, our study retrieves 43 novel Fkh1 targets and 38 novel Fkh2 targets (Fig. 3.4), among which 3 and 6, respectively, are enzymes (see Table 1 and Supplementary Excel Table S6).

The correlation between Fkh and target expression level

To evaluate the quality of our results, we monitored the correlation between the expression level of Fkh1 and Fkh2 and their targets, by using publicly available gene expression datasets. We combined the target genes identified in exponential and stationary phases (listed in Supplementary Excel Table S2) for each Fkh transcription factor, and analyzed them using the SCEPTRANS database (`http://moment.utmb.edu/cgi-bin/sceptrans.cgi`) [42]. We tabulated the total number of genes and the number of retrieved target genes that are correlated with Fkh1 or Fkh2, based on correlation coefficient thresholds of 0.60 and 0.80, across the nine microarray datasets from five studies in SCEPTRANS [8, 43–46]. The correlated genes grouped by the threshold and by microarray dataset are listed in Supplementary Excel Table S7. In total, 305 and 157 of the retrieved target genes (72% and 69%) correlate in terms of expression level with Fkh1 and Fkh2 expression, respectively, in at least one of the nine datasets.

Furthermore, we tabulated the genome-wide fraction of genes correlated with Fkh1 and Fkh2 across each of the nine datasets. By multiplying that fraction

with the number of target genes, we calculated the expected number of correlated ChIP-exo target genes if the target genes were randomly selected from the total pool of genes. We then calculated the ratio of the actual number of correlated ChIP-exo target genes and the expected number. We observed an enrichment in correlated Fkh1 target genes (i.e. a ratio > 1.5) in eight out of nine microarray datasets. Fkh2-correlated target genes were enriched in six out of eight microarray datasets (see Supplementary Excel Table S7).

Dynamics of cell cycle-regulated target genes highlight a distinct activation of Fkh1 and Fkh2 functions across cell cycle phases

An earlier study applied a deconvolution algorithm to one of the nine microarray datasets analyzed above (43) and has identified 1,082 genes as being cell cycle-regulated (i.e., expressed cyclically), among which 198 metabolic enzymes, reporting the time of peak expression and cell cycle phase where it occurs for each such gene [18]. Fkh1 and Fkh2 were considered part of the 'high-quality' set of 694 cell cycle transcriptionally regulated (CCTR) genes with 95% confidence or better. Subsets of 84 and 122 target genes for Fkh1 and 46 and 65 target genes for Fkh2 belong to the extended CCTR set for exponential and stationary phases, respectively (see Table 3.1). The main expression peaks of Fkh1 and Fkh2 were identified to occur at 67 and 3 minutes during S and G1(P) phase, respectively. In addition, Fkh2 did exhibit a secondary expression peak at 74.5 minutes during S phase.

We have analyzed the subset of identified targets that are cell cycle-regulated (Supplementary Excel Table S8) in terms of their cell cycle phase of peak expression (Fig. 3.5). When comparing the distributions of the identified target genes in the four ChIP-exo experiments to the genome-wide distribution [18], for both Fkh1 and Fkh2 we observed an enrichment of targets whose expression peaks in the mid cell cycle (S phase) and an underrepresentation of targets that peak in the early cell cycle (G1, G1(P), G1/S phases), in both exponential and stationary phases (see Fig. 3.5 and Table 3.2). The enrichment of targets that peak in S phase is somewhat, yet significantly higher for Fkh1 than Fkh2. Conversely, Fkh2 but not Fkh1 targets are enriched in the late cell cycle (G2, G2/M, M, M/G1 phases), consistent with earlier data showing that Fkh2 is expressed during the late stages of the cell cycle [19]. Analyzing the data in more detail, we observed that both Fkh1 and Fkh2 targets are shifted towards S and G2 and away from G1, G1(P), G1/S and M, in both exponential and stationary phases. Moreover, Fkh1 and Fkh2 targets show an opposite trend at the G2/M and M/G1 transitions as compared to the genome-wide distribution [18]: Fkh1 targets are underrepresented in G2/M and enriched in M/G1, whereas Fkh2 targets are enriched in G2/M and underrepresented in M/G1. Taken together, these findings highlight a tendency for the Fkh1 targets to peak earlier (in the S phase) as compared to the Fkh2 targets, which peak in the late cell cycle phases (G2 through M/G1 phases). Remarkably, these tendencies are small however. This may explained by the fact that we here only look at the location of the peak of expression which although indicative cannot provide a full picture. Further analysis into the complete time

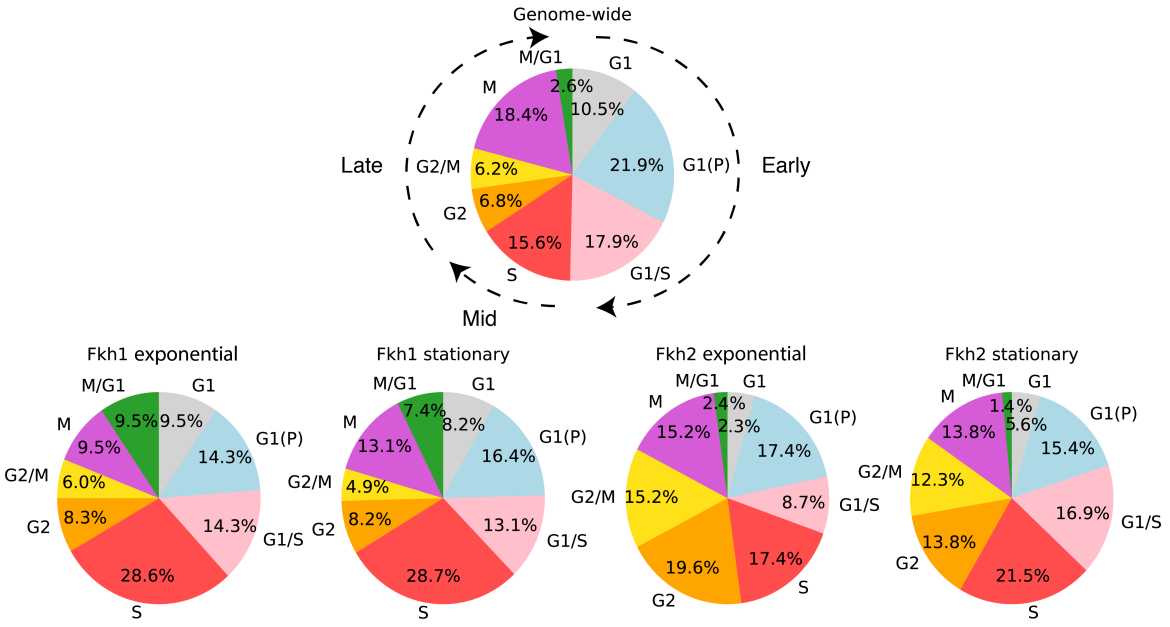


Figure 3.5: Distribution of the phases of peak expression for the mRNA of cell cycle-regulated Fkh target genes. A genome-wide dataset [18] was compared to the ChIP-exo Fkh dataset to identify target genes that are cell cycle-regulated. The distribution of the cell cycle-regulated genes [18] is shown on top, whereas the other four pie charts show the distribution of Fkh1 and Fkh2 targets, in both exponential and stationary phases.

Condition	Early (G1, G1(P), G1/S)	Mid (S)	Late (G2, G2/M, M, M/G1)
Fkh1 exponential	-12%	+13%	-1%
Fkh1 stationary	-13%	+13%	0%
Fkh2 exponential	-20%	+2%	+18%
Fkh2 stationary	-13%	+6%	+8%

Table 3.2: Enrichment of targets for which mRNA expression peaks in the early, middle or late cell cycle phases as compared to the cell cycle-regulated genes as previously identified [18]. The percentages reported are inferred from the pie charts shown in Fig. 3.5. The percentages for each condition are calculated as the difference with respect to the genome-wide dataset.

course data from [18] may provide a fuller picture and is likely to complement this view.

Fig. 3.6 visualizes the set of cell cycle-regulated targets of Fkh1 and Fkh2 in exponential phase as a stack plot. Each target is coloured according to the function associated to the GO annotation, which was performed through GEMMER (see Materials and Methods). The stack plot of Fkh1 and Fkh2 targets in stationary phase is visualized in Supplementary Text, Fig. S3.8. The position on the y-axis within each column - corresponding to a cell cycle phase where the expression is maximal for each gene - is dictated by the *maxPeak* SNR of the ChIP-exo experiments. We observe that the majority of Fkh1 cell-cycle regulated targets

show their expression peak in the S phase. Moreover, when focusing on the cell cycle-regulated enzymes (indicated in bold in Fig. 3.6A) across different cell cycle phases, we observe that among the 24 Fkh1 enzymatic targets the majority is enriched in the early and mid cell cycle (G1, G1(P), G1/S and S phases) as compared to the late cell cycle (G2, G2/M, M and M/G1 phases). Conversely, the 8 enzymatic targets of Fkh2 are equally distributed throughout early and late cell cycle phases (indicated in bold in Fig. 3.6B). These findings suggest that Fkh1 cellular functions, mediated by the activity of its targets, are realized earlier than Fkh2 functions, which do not seem to be confined to a specific cell cycle phase.

Using the *CDC28* data from [18] as an informative example (see Supplementary Text), target genes that are cell cycle regulated, with expression peaks within a window of -25 to 45 minutes after Fkh1 and Fkh2 expression peaks, may be considered to align with expected behavior for Fkh1- and Fkh2-regulated genes. This implies a target window of 42 – 102 minutes (i.e. from the end of G1/S to mid G2 phase) for Fkh1. For Fkh2, this would suggests two target windows: from (i) 278 – 48 minutes (from the end of G1 to the start of G1/S phase) and from (ii) 50 – 110 minutes (i.e. from the end of G1 to mid G2 phase). We conclude that genes listed in Fig. 3.6 and Fig. S3.8 that fall within these time windows show expected behavior for genes regulated by Fkh1 and Fkh2. The well-known Fkh1/Fkh2 target genes *CLB1* and *CLB2* occur within these windows; furthermore, *CLB3*, which we have shown to be regulated by Fkh2 [16] falls within the Fkh2 window. However, it should be noted that, given that Fkh2 exhibits two expression peaks, it may well be present in the intermittent time-period as well so that targets peaking in the window 110 – 278 should not be discounted.

Functional enrichment of identified Fkh target genes

For all identified Fkh target genes we performed an overrepresentation analysis for GO terms with respect to the biological processes they are involved in, by using Fisher's exact test through the PANTHER database [47]. We found several significantly overrepresented terms for a False Discovery Rate (FDR) threshold of 0.05 (Supplementary Excel Table S9 lists the FDR for all GO terms across all experimental conditions). The GO terms for *cell cycle* and *mitotic cell cycle* were enriched across all four ChiP-exo experiments. Furthermore, the GO terms for *(mitotic) cell cycle* and *cell division* were enriched across three out of four ChiP-exo experiments (lacking in Fkh2 stationary phase and Fkh1 stationary phase, respectively). Fkh1 has uniquely enriched terms for *organelle fission*, *(mitotic) nuclear division* and *(mitotic) sister chromatid segregation*. Moreover, the Fkh1 exponential experiment showed a unique enrichment in the terms for *(nuclear) chromosome segregation*, whereas the Fkh1 stationary experiment showed unique enrichment in the terms for *regulation of cell cycle* and *regulation of (mitotic) cell cycle process*. Finally, the Fkh2 stationary experiment showed no uniquely enriched terms, whereas the Fkh2 exponential experiment showed a unique enrichment in the terms *(fungal-type) cell wall organization* and *external encapsulating structure organization*. In addition to the formal enrichment test of GO terms, the Fkh target genes identified in the four ChIP-exo experiments were analyzed for their global,

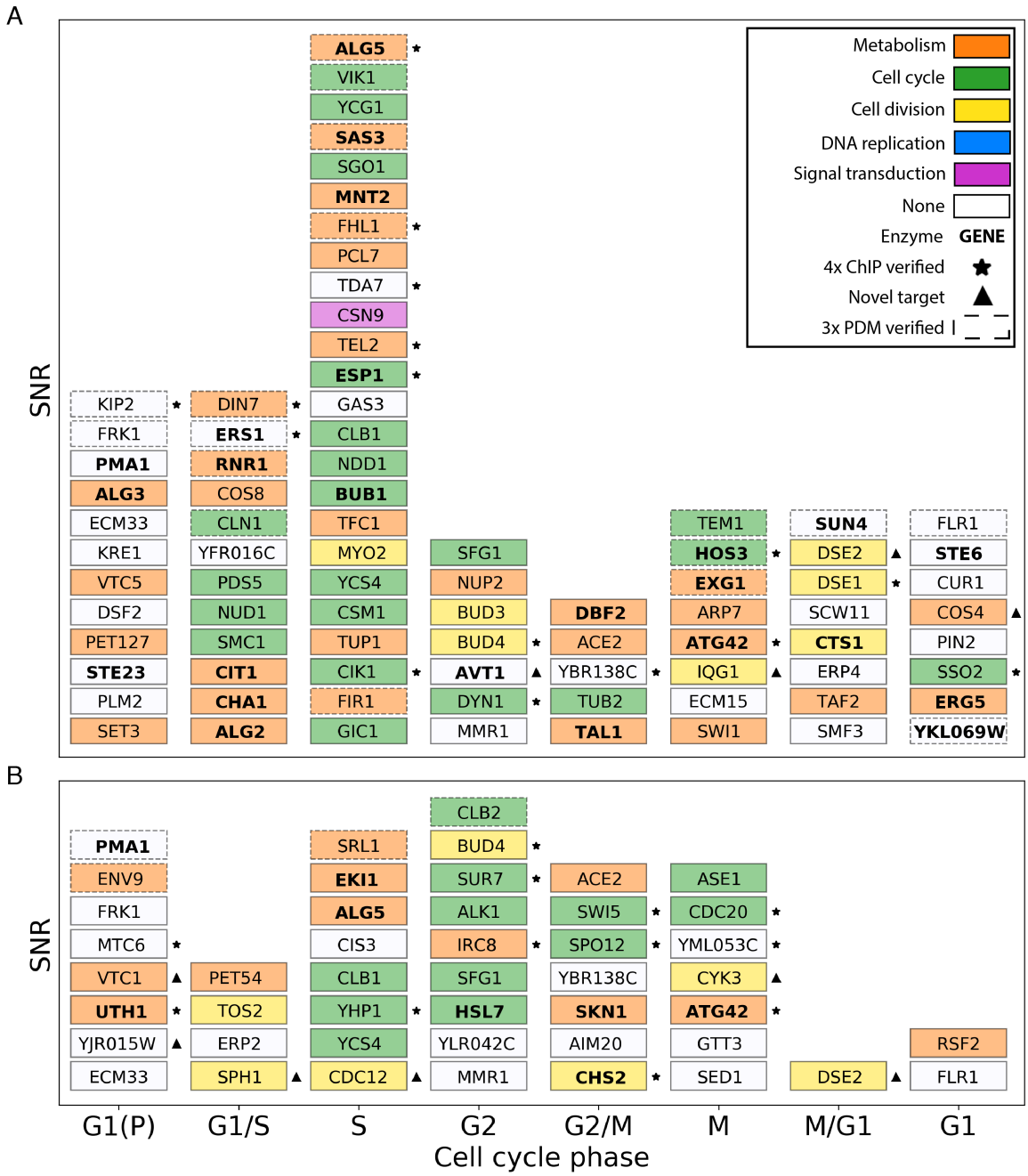


Figure 3.6: Stack plot of target genes identified by ChIP-exo in exponential phase that have a cell cycle-regulated peak expression level [18]. (**A**) Fkh1 target genes. (**B**) Fkh2 target genes. Within each column a higher position on the y-axis indicates a higher *maxPeak* SNR. The x-axis indicates the phases of peak expression as reported [18]. The color for each target gene indicates its major biological function if identified in GEMMER [31]. Targets marked with an asterisk are verified by all four (4x) ChIP studies, whereas targets marked with a triangle indicate novel target genes that have not been reported in the previous ChIP studies. Targets identified as significant by all three PDMs are shown with a dashed border.

rather than for their specific, function, showing an enrichment of targets with a function in cell cycle and cell division (see Supplementary Text, Table S4). This result supports the earlier finding that Fkh targets are primarily cell cycle genes [8].

Interestingly, even though no GO terms related to metabolism were enriched in the analyses above, we observed that a fraction of genes with a metabolic function was present among the Fkh targets. However, statistically significant enrichment of the GO terms is not required for functional impact of Fkh1 and Fkh2 on metabolic processes. Specifically, we identified 60 and 18 enzymatic targets of Fkh1 and Fkh2, respectively, in exponential phase, and 103 and 50 enzymatic targets, respectively, in stationary phase, most of which catalyze metabolic reactions (see Table 3.1 and Supplementary Excel Table S10). This provides a clear indication of the potential role of Fkh1,2 as hubs connecting cell cycle and metabolism.

Fkh targets in their functional context through projection onto KEGG Pathways

With the aim to explore the pathways where a metabolic function was observed for Fkh targets, our ChIP-exo results were superimposed on a set of 25 KEGG maps of interest, in order to intuitively display the (metabolic) function of Fkh targets (see Materials and Methods), by using the Pathview library for R (see Supplementary KEGG Figures). In particular, we focused on Fkh1 and Fkh2 targets in central carbon metabolism as identified by ChIP-exo (Fig. 3.7). In Supplementary Text, Fig. S3.9 a similar overview includes Fkh targets previously identified in ChIP-chip studies that were not recovered by ChIP-exo. Noteworthy, 16 (iso)enzymes catalyzing 14 reactions in the visualized part of central carbon metabolism are potentially regulated by Fkh. 14 enzymes out of 16 are potential Fkh1 targets and 5 enzymes out of 16 are potential Fkh2 targets, pointing once again to a predominant metabolic role for Fkh1 as compared to Fkh2. Remarkably, all three isoenzymes of the citrate synthase (*CIT*), the entry point enzyme of the TCA cycle, as well as enzymes involved in ethanol fermentation from pyruvate were retrieved as targets. In detail, for the TCA cycle: *CIT1* as Fkh1 target in both exponential and stationary phases; *CIT2* as Fkh1 target in stationary phase; and *CIT3* as both Fkh1 and Fkh2 target in exponential phase. For the ethanol fermentation, the pyruvate decarboxylase *PDC1* and the alcohol dehydrogenase *ADH4* were retrieved as Fkh1 targets in both exponential and stationary phases. All 16 enzymatic targets, with the exception of *GND1* and *IDH2* for Fkh2 and *CIT3* for Fkh1, have been previously reported by ChIP-chip studies [24, 25].

We annotated the ChIP-exo dataset with the KEGG Pathways that each of the 7,217 target genes occurs in. Together, the Fkh targets in all four experimental conditions map onto 89 distinct KEGG pathways, ranging from cell cycle to signaling and metabolism (see Supplementary Excel Table S11). In Supplementary Excel Table S10, all enzymatic targets of Fkh1 and Fkh2 in exponential and stationary phases are reported with the KEGG Pathways they occur in and their cell cycle phase of peak expression (if available). Moreover, examples of Fkh1,2 targets in autophagy, signal transduction and cell cycle are shown in Supplemen-

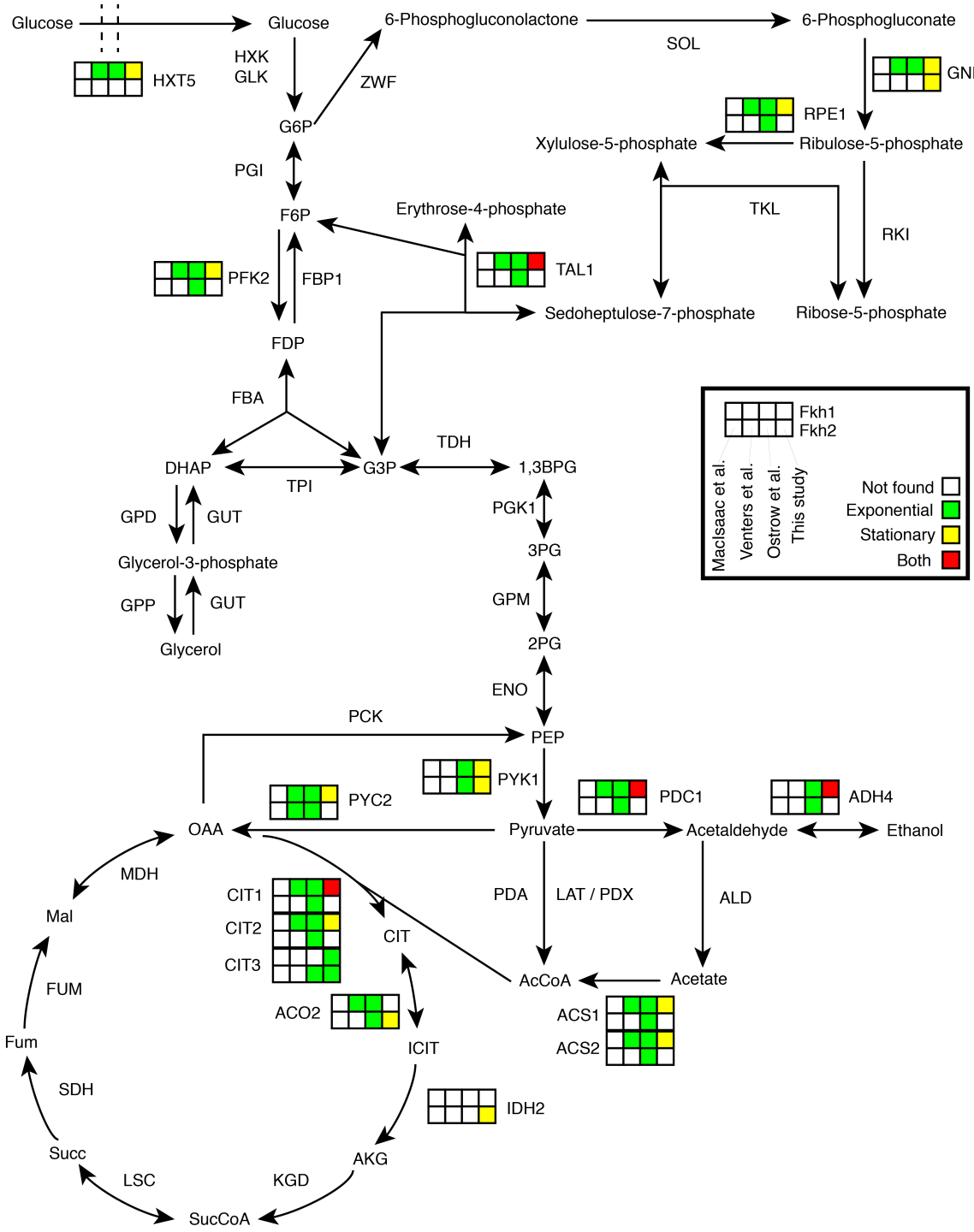


Figure 3.7: Overview of metabolic enzymes in central carbon metabolism that are targets of Fkh1 and Fkh2. Each enzyme is associated with eight squares divided in two rows (Fkh1, top row; Fkh2, bottom row) representing data analysis of four different genome-wide studies: MacIsaac et al., Venters et al., Ostrow et al. and this study. Empty squares indicate genes that were not retrieved as significant targets, whereas colored squares indicated a positive evidence. A distinction between the results in exponential and stationary phases is visualized through the color of the squares (see the figure insert). Isoenzymes that have no available evi-

Figure 3.7: (Continued) dence in any of the three studies were neglected. In some cases, metabolic enzymes may have no associated squares when no isoenzyme is available with an experimental validation.

tary Text, Figure S10. To illustrate the multi-scale nature of Fkh1 and Fkh2 target genes, we highlighted 13 pathways and the Fkh target genes that function therein in Supplementary Text, Table S3.5 .

Altogether, our findings provide the field with an up-to-date overview of the current knowledge of Fkh targets within their functional context.

3.4 Discussion

As compared to previous ChIP-based methodologies, our ChIP-exo analyses were performed for the Fkh1 and Fkh2 transcription factors in two experimental conditions: exponential and stationary phases. Due to a relatively high number of targets identified by the GEM and MACE PDMs we have developed a novel PDM that we named *maxPeak* and used it alongside GEM and MACE for the analysis of our ChIP-exo dataset. We considered only those target genes as targets that scored above threshold (see Fig. 3.1 and Supplementary Text, Fig. S3.2–S3.4) in at least two out of three methods. In this way we aimed at minimizing the occurrence of false positives and false negatives and maximizing true negatives and true positives. By basing the *maxPeak* threshold on the minimum score attained by target genes that were predicted by both GEM and MACE given their respective thresholds, the *maxPeak* score essentially forms the deciding factor that decides whether genes that were only predicted by either GEM *or* MACE should actually be considered targets. The more intuitive nature of the *maxPeak* technique, which looks only at the highest signal in an upstream region of the TSS of a gene, and compares the set of these signals genome-wide, thus provides a sanity check on the genes for which the predictions by GEM and MACE disagree. Given the correlations between the Fkh and their respective target genes identified using our methodology (see main text) we feel that our approach has worked satisfactorily.

We have observed that different ChIP-based methodologies retrieve different numbers and collections of Fkh targets. However, by analysing the verified Fkh targets in common between various studies, we have provided a comprehensive view of the most likely genes whose expression may be modulated by Fkh1 and Fkh2. Noteworthy, when we conducted the data analysis comparing the outcome of the three PDMs (*maxPeak*, GEM and MACE) we observed that the divergence in the target genes retrieved between GEM and MACE as well as between both software tools and the novel *maxPeak* method is substantial (a detailed analysis is presented in Supplementary Text, Fig. S3.5–S3.6). One reason for these different predictions is highlighted in Fig. S3.1: the algorithms differ substantially and may pick up different features of DNA regions that show binding. This evidence points out a need in the field to (i) identify the stability of the methods, (ii)

investigate advantages and shortcomings of each method, and (iii) measure the accuracy of the methods with regard to the identification of functional targets.

Our analysis of the targets retrieved by at least two out of three PDMs solidifies and enriches the global perspective on the functions possibly exerted by Fkh. First of all, there are sets of targets in common among multiple ChIP-based studies (Fig. 3.4) that reinforce the role of Fkh in the cell cycle and point towards possible roles in signal transduction and metabolism (see Fig. 3.4 and accompanying text). Second, Fkh targets that have been identified as cell cycle-regulated peak in expression level across all phases of the cell cycle but are enriched in S phase and G2, M and M/G1 phase for Fkh1 and Fkh2 respectively (see Fig. 3.5). The spread in timing of the expression peaks of the possible targets together with the wide-array of functionalities of these target genes (see Fig. 3.6) hints at broad functionalities for both Fkh. Third, even though the enrichment of metabolic genes was not statistically significant, the high number of target genes with functions in metabolism and the 89 KEGG pathways affected by their targets (See Fig. 3.7 and Table S3.5), point to hitherto unknown roles of the Fkh. Together, our analysis highlights the role that Fkh have as hubs that integrate multi-scale regulatory networks, exemplified by metabolism and cell cycle, to achieve proper timing of cell division.

Fkh1 and Fkh2 are paralogs that have diverged, with a protein identity of 71% and a protein similarity of 85% [48]. The previously identified canonical Fkh1/Fkh2 binding motif 5'-GTAAACAA-3' reported in the YeTFaSCo database [27] and by MacIsaac et al. [26], is present in over 1400 locations throughout the genome. To analyze the enrichment of this binding motif on our ChIP-exo dataset, we extracted all peak locations corresponding to the target genes spanning a -250 to +250 bp window around the peak location. We combined the sequences for both exponential and stationary phase experiments for each transcription factor to obtain a robust motif identification. This collection of sequences was analyzed using three algorithms from the MEME-suit [49] with complementary characteristics: MEME [50], DREME [51] and CentriMo [52]. MEME and DREME identify long and short ungapped motifs, respectively, whereas CentriMo identifies known DNA-binding motifs from other transcription factors. The top significantly enriched motifs returned by MEME and DREME either contain, or are similar to, the canonical DNA-binding motif (Fig. 3.8). The DREME motifs for Fkh1/Fkh2 are virtually equal to the canonical motif, and differ among each other only in terms of possible alternatively preferred bases at 2 or 3 locations. The enriched sequence pattern identified by MEME is longer than the established canonical motif (19-21 bp) but the latter can be clearly identified within it for both Fkh1 and Fkh2. We observe that this top motif is very similar for both Fkh1 and Fkh2, but with a changed and increased preference for the surrounding bases for Fkh2. The presence of a longer and more specific motif for Fkh2 as compared to Fkh1 may translate to a different set of target genes and/or a different affinity for such target genes.

Given the similarity of their protein sequence and DNA-binding motifs, we were interested in exploring the overlap between targets of Fkh1 and Fkh2 in both exponential and stationary phases. The ChIP-chip studies [24–26] already

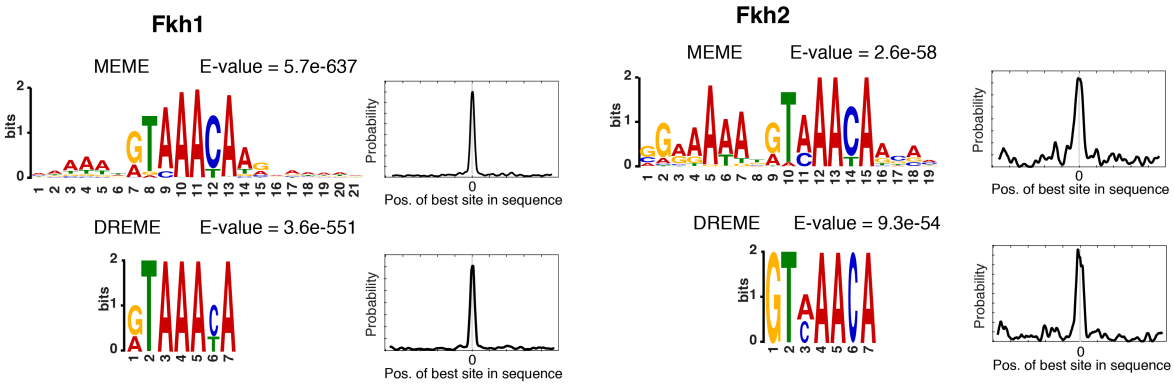


Figure 3.8: Top DNA-binding motifs for Fkh1 and Fkh2 based on the ChIP-exo target gene peak sequences, as identified by the MEME and DREME algorithms using MEME-ChIP.

Targets genes	exponential	Stationary	4x ChIP verified
Overlap	26	65	3
Fkh1 specific	265	351	26
Fkh2 specific	79	155	12

Table 3.3: Overlap between Fkh1 and Fkh2 target genes. The columns with exponential and stationary data refer to this study. The 4x verified ChIP column refers to the experiments in exponential phase performed in this study, MacIsaac et al. (MacIsaac et al., 2006), Venters et al. (Venters et al., 2011) and Ostrow et al. (Ostrow et al., 2014).

showed a large set of unique targets, with only 11% – 44.1% of identified Fkh1 targets shared with Fkh2 (see Supplementary Text Table S3.6). In Table 3.3 we report the number of overlapping ChIP-exo Fkh targets in the two experimental conditions, as well as among the set of 4x ChIP verified targets. Using ChIP-exo we observed fewer common targets than in the published ChIP-chip studies, strengthening the hypothesis of divergent functions for Fkh1 and Fkh2. The percentage of overlapping Fkh target genes is 7% and 11% for exponential and stationary phases, respectively. Considering the number of overlapping versus specific (Fkh1 only and Fkh2 only) targets reported in Table 3.3 , we conclude that the vast majority of Fkh targets is unique for Fkh1 or Fkh2 specific functions. A similar outcome was observed for the 4x ChIP verified targets, with a percentage of overlapping Fkh target genes equal to 7%. These data suggest that, regardless of the different ChIP methodologies employed, Fkh1 and Fkh2 appear to have divergent functions.

The observation of divergent target genes for Fkh1,2 is further highlighted by the suggestion of a potential metabolic function for both Fkh1 and Fkh2 due to the presence of metabolic enzymes among the target genes (Table S3.1, S3.5 and Fig. 3.7, S3.8 and S3.9). Most of these targets are not in common and their activity is realized at different times throughout cell cycle regulation (3.6). The analyses

of the subset of cell cycle-regulated targets (Table 3.2 and Fig. 3.5 and 3.6) indicates target genes peaking in expression across the entire cell cycle. However, we observe a major role for Fkh1 in the early cell cycle (from G1/S through S phases) (Fig. 3.6) with an enrichment of target genes in S phase specifically (Table 3.2). In contrast, we observe a more equally distributed role for Fkh2 (Fig. 3.6) with an enrichment of targets active in the late cell cycle (G2 through M/G1 phases) (Table 3.2). These findings are in agreement with early data showing that Fkh1 is expressed earlier than Fkh2 [19].

We investigated the height of the ChIP-exo signal upstream of non-overlapping target genes for the Fkh for which they were not considered a target. We observed that roughly half (44% – 59%) of the genes that we list as unique target genes in Table 3.3 cross the threshold in one of the three PDMs and that, similarly, the other half does so in none of the PDMs. The latter subset supports the conclusion that there are substantial differences in the set of target genes. Simultaneously, the former subset points out that many of the unique target genes may show some limited, but lower binding affinity, for the other Fkh, potentially indicating a compensatory interplay between the two transcription factors. This result calls for a detailed investigation of the relative binding affinities of targets shared between Fkh1 and Fkh2.

It remains speculative whether or not the differences in the observed motifs for Fkh1 and Fkh2 contribute to the difference in target genes that we retrieved. It has previously been observed that both redundant and different functions for Fkh1 and Fkh2 exist, and that these differences were not attributable to the DNA-binding domain [14]. Fkh1 and Fkh2 are paralogs with a similar DNA-binding motif, and we observed that they bind only to partially overlapping sets of a target genes (potentially with a different affinity). This evidence suggests that the evolutionary divergence between the two transcription factors, together with the shift in the timing of the expression window [18], left in place a common set of redundant functions but, over time, gave rise to more specific sets of target genes.

Aside from a difference in the main binding motif, it is possible that that Fkh1 and Fkh2 bind different secondary motifs or interact (in complex) with different secondary transcription factors. Our motif analysis suggested secondary enriched motifs which differed between Fkh1 and Fkh2 that are similar to binding sequences of other transcription factors. Mcm1 acts as a scaffold protein for both Fkh2 and the co-activator Ndd1, regulating the G2/M transition and, thus, cell division [10, 12]. Since the Mcm1 motif showed an E-value of 10^{-12} for Fkh2, we considered all enriched motifs above this threshold (the E-value of a motif estimates the number of motifs that would have equal or higher log likelihood ratio if the input sequences had been generated randomly). We found enriched motifs matching eight transcription factors (Ecm22, Azf1, Ixr1 Hmlalpha2, Mcm1, Hmra2, Matalpha2 and Dal82) for Fkh1 and two transcription factors for Fkh2 (Gcn4 and Mcm1). We observed that the Azf1 and Hmra2 motifs were very similar to the canonical Fkh motif, therefore disregarded these. None of the remaining transcription factors with similar binding motifs have known physical or genetic interactions with Fkh1 or Fkh2, with the exception of Mcm1. For several transcription factors with similar binding motifs of Fkh1, genetic evidence of an in-

teraction with Fkh2 is available. Specifically, a genetic interaction of Fkh2 was suggested with Ixr1 and Rox1 [53], transcriptional repressors that regulate hypoxic genes during normoxia. Furthermore, a genetic interaction was reported between Fkh2 and Dal82 [54], regulator of allophanate inducible genes. We currently envision no reason for the enrichment of these motifs within peak regions. However, this evidence call for detailed experimental investigations of the possible interplay between Fkh1 and Fkh2 and these transcription factors.

Interestingly, our work highlights a number of metabolic enzymes as targets of Fkh1 and Fkh2, 16 of which play a role in central carbon metabolism (Fig. 3.6 and 3.7): *HXT5*, *GND1*, *RPE1*, *TAL1*, *PFK2*, *PYK1 / CDC19*, *PYC2*, *PDC1*, *ADH4*, *CIT1*, *CIT2*, *CIT3*, *ACS1*, *ACS2* (Fkh1 targets) and *GND1*, *PYK1 / CDC19*, *CIT3*, *ACO2*, *IDH2* (Fkh2 targets). The deletion of two of these enzymes is lethal: pyruvate kinase (*PYK1 / CDC19*) targets of both Fkh1 and Fkh2 and acetyl-coA synthetase (*ACS2*) target of Fkh1. Furthermore, deletion of many among the other 16 genes results in reduced growth rates in a number of experimental conditions (e.g. *GND1*, *RPE1*, *PFK2*, *PYC2*, *PDC1*, *CIT1*, *CIT2*, *ACS1*, *ACS2* and *IDH2*). Consequently, the altered growth rate observed in *fkh1*Δ, *fkh2*Δ and *fkh1*Δ *fkh2*Δ mutants [14] may be due to absence in the regulation of one or more of the 16 target enzymes in the central carbon metabolism. Thus, our work suggests a potential route for a role for both Fkh1 and Fkh2 in central carbon metabolism.

When focusing on the 24 (Fkh1) and 8 (Fkh2) cell cycle-regulated metabolic enzymes across different cell cycle phases (indicated in bold in Fig. 3.6 and Supplementary Text, Fig. S3.8), we observed that Fkh1 and Fkh2 target several cyclically expressed metabolic enzymes that are involved in membrane processes, which are centred around the two major cell cycle transitions. Specifically, Fkh1 targets *PMA1*, *ALG3*, *ERS1*, *ALG2*, *ALG5* and *MNT2* around the G1/S transition (G1(P) – G1/S – S) and *EXG1*, *SUN4* and *STE6* around the M/G1 transition (M – M/G1 – G1). Similarly, Fkh2 targets *PMA1*, *UTH1* and *ALG5* around the G1/S transition (G1(P) – G1/S – S). We hypothesize that Fkh transcription factors, which are not active after cell cycle exit until the next S phase upon Ndd1 activation [55], affect the plasma membrane by: (i) switching on their targets centred around the G1/S transition, and (ii) switching these off in S phase due to Fkh activation upon binding of the co-activator Ndd1 [9, 12]. Following this line of thought, Fkh1 will subsequently affect the plasma membrane at the late cell cycle phases, when Fkh are inactivated until the following S phase.

The Forkhead family of transcription factors is defined by a shared DNA-binding motif, referred to as the winged-helix domain. The mammalian Forkhead family encompasses 18 subfamilies [56, 57], and the human genome contains over 40 FOX genes. Of these, the FoxM1 and FoxP proteins represent the closest homologs of Fkh1,2 [48, 58]. FoxM1 was identified to have the in vitro DNA-binding consensus site TAAACA [59]. This motif shares the core sequence recognized by other members of the Forkhead family [59] and also matches part of the motif others [26] and we, identified for Fkh1,2. The similarity in the binding motif suggests that some of the target genes retrieved in this work may carry over to the FoxM1 transcription factor. FoxM1 is involved in cell cycle regulation, stress response, chromatin silencing, and aging [58]. However, if the suggested Fkh1,2-mediated

regulation of metabolic genes would translate to FoxM1, it would be especially interesting since FoxM1 has already been implicated in cell division by regulating the expression of the mitotic Cyclin B [60], homolog of Clb2, and its expression has been observed in multiple tumor-derived cell-lines (see [61] and references therein).

Considering that our data point to multiple roles for Fkhs in cell cycle progression, we have explored the relevance of our ChIP-exo findings for Fkh1 and Fkh2 effects on cell cycle dynamics (Fig. 3.9). In Fig. 3.9A the regulatory cascade driving phase-specific events in cell cycle progression is shown: in G1 phase, the cyclin Cln2, together with the kinase Cdk1, inhibits the cyclin/Cdk1 inhibitor Sic1. When Sic1 activity is blocked, Sic1-mediated inhibition of Clb5/Cdk1 is released, allowing it to activate substrates required for DNA replication in S phase. Subsequently, a Clb/Cdk1 cascade is activated, involving waves of Clb5, Clb3 and Clb2 cyclins (all bound to Cdk1). These waves of cyclins are responsible for the control of DNA replication and mitotic entry/exit from S through M [62, 63]. In Fig. 3.9B we summarize the evidence of Fkh binding at promoters of target genes in this cascade. In our ChIP-exo study, *CLB2* is confirmed to be a major target of Fkh1 and Fkh2, as reported [9–15]. Our findings highlight that this is evident in both exponential and stationary phases for Fkh2. We also confirm *SWI5* being target of Fkh2 [10, 11, 64] and *FKH2* being target of Fkh1, as reported by multiple genome-wide studies [24, 25].

Besides known verified targets in the cell cycle cascade, our ChIP-exo experiments highlight three Fkh targets in the cell cycle cascade, previously identified only by some but not all the ChIP-chip studies, for which experimental validation is currently lacking: *CLN1* and *ACE2* being targets of both Fkh1 and Fkh2, and *NDD1* being a target of Fkh1 (not shown). Further analyses are required in order to validate these findings. If validated they will shed light on possible novel regulatory mechanisms of Fkhs in cell cycle regulation.

Our study also points to limitations of genome-wide studies, including ChIP-exo, in the identification of targets, such as the *CLB3* gene. Fkh2 binding to *CLB3* promoter was shown only by one ChIP-chip study [24]. Furthermore, we have recently demonstrated that Fkh2 binds to the *CLB3* promoter and regulates Clb3 expression, thus synchronizing the temporal expression of mitotic *CLB* genes in a linear cascade (Clb5 - Clb3 - Clb2) [16]. However, in our ChIP-exo data for Fkh2, *CLB3* does not score above threshold in any of the three PDMs, and therefore it was not considered as a Fkh2 target gene (false negative). This example highlights that genes that show low DNA binding signal in ChIP studies should not be regarded as *not* being regulated. Conversely, a potential regulation may be suggested for high-scoring target genes. Binding data of transcription factors provide an indication of potential regulatory activities; however, these are not proof of such activity, for which an experimental validation would be required.

In addition, our findings do not support previously suggested Fkh targets: *SWI4* and *SWI6* for both Fkh1 and Fkh2 (Venters et al., 2011), *SIC1* for Fkh1 [24, 25], and *CLB5* for Fkh2 [24]. The latter scenario has been recently excluded by our independent experimental analyses, showing that *CLB5* may not be a Fkh2 target [16], thus highlighting the occurrence of false positives identified by previous

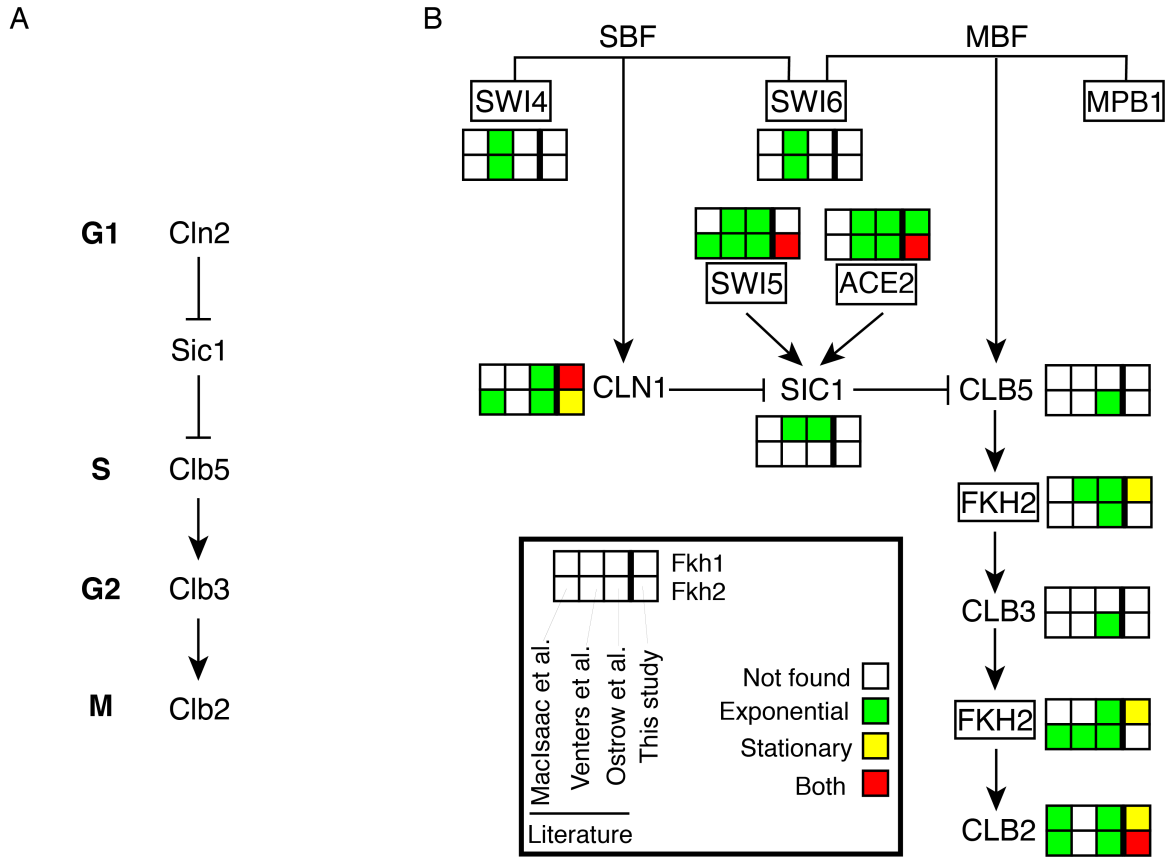


Figure 3.9: Fkh1 and Fkh2 target genes in the molecular cascade regulating dynamics of cell cycle progression. (**A**) Molecular players driving phase-specific cell cycle events (see text for details). (**B**) Overview of cell cycle regulators that are Fkh targets. The transcription factors *SWI4*, *SWI6*, *MPB1*, *SWI5*, *ACE2* and *FKH2* are shown within rectangles.

ChIP-chip studies.

We observed a higher number of correlated retrieved target genes than randomly expected for both Fkh1 and Fkh2 across nine publicly available microarray datasets (Supplementary Excel Table S7). This work points towards future studies aimed to the experimental validation of the targets retrieved, by assessing changes in gene expression upon Fkh knockout. In our view, priority should be given to: (i) high scoring target genes as ranked by all three PDMs (Supplementary Excel Table S3), (ii) high scoring target genes identified by all four ChIP studies available for Fkh1 and Fkh2 (see Fig. 3.6 and Supplementary Excel Table S5), and (iii) target genes that we identified as highly correlated with Fkh1 and Fkh2 in available gene expression studies (Supplementary Excel Table S7).

Finally, by referencing the KEGG pathways the target genes mapped on, and providing the number of metabolic targets, our analyses highlighted the potential of Fkh1 and Fkh2 to connect their specific functions within the core cell cycle network with other regulatory processes in metabolism and signal transduction. Altogether, the data presented in this study clearly provide evidence of the wide-reaching influence of Fkh, and open avenues for further research by pointing to

the Fkh transcription factors as hubs that integrate multi-scale regulatory networks to achieve proper timing of cell division in budding yeast.

Data Availability

A collection of Python scripts reproducing the data integration and Jupyter notebooks reproducing the data analysis are available as Supplementary Code Repository and as part of a Github repository (`https://github.com/barberislab/ChIP-exo_Fkh1_Fkh2`).

Supplementary data, code and tables are also available at NAR online: http://doi.org/10.1093/nar/gkz603.

Supplementary Materials and Methods

Data annotation through GEMMER

GEMMER is a novel web-based data-integration and visualization tool that we recently developed for budding yeast [31] (and see Ch. 4. GEMMER integrates protein-coding genes, interactions and general and functional annotation from the Saccharomyces Genome Database (SGD) [65], localization and abundance data from CYCLoPs [66] and YeastGFP [67, 68] databases, and both the time and cell cycle phase of peak occurrence of RNA transcript levels [18, 42]. We merged the processed ChIP-exo data for all genes with their annotations by querying the GEMMER SQL database and merging the information into a new dataset. In addition, information about literature studies that report a gene as a target, based on ChIP-chip data, was included [24–26]. The MacIsaac et al. data [26] was retrieved through GEMMER. The Ostrow et al. data [24] was retrieved from their Supplementary Excel file S4. The Venters et al. data [25] was retrieved from their 25C_UTmax Table S4a, using a 5% FDR threshold to select targets of Fkh1 and Fkh2. The Supplementary Code Repository contains Python scripts that reproduce this analysis.

Conceptual comparison between GEM, MACE and *maxPeak*

We performed peak detection by three different methods: *maxPeak*, GEM and MACE, and compared the sets of significant target genes retrieved by all of these methods. Fig. S3.1 conceptually summarizes the principles behind the three peak detection methods (PDMs) and their specific differences.

Setting thresholds for GEM, MACE and *maxPeak*

To set thresholds that define which genes are considered targets by each of the three PDMs, we generated three 2x2 score comparisons (see Fig. S3.2 – S3.4). We observed a linear trend between GEM and *maxPeak* SNRs (Fig. S3.2), as well as a boundary (p-value ~ 0.005) beyond which MACE p-values seem to be enriched (Fig. S3.3 and S3.4). We chose to use the GEM threshold of SNR ≥ 1 and set the MACE threshold to p-value ≤ 0.005.

We considered any target gene that is retrieved as significant by both GEM and MACE as a confident target and therefore we set the threshold of significance for *maxPeak* to the lowest score obtained across all four experimental conditions by any gene that was retrieved by both GEM and MACE. This turned out to a threshold for *maxPeak* of SNR ≥ 1 (rounded down from 1.07). Consequently, in Fig. S3.3) genes with circles in the upper-right quadrant of each panel were considered targets together with blue circles in the upper-left and bottom-right quadrants of each panel. Similarly, for Fig. Fig. S3.4) and Fig. S3.5) genes with circles in the upper-left quadrant of each panel were considered targets together with blue circles in the bottom-left and upper-right quadrants of each panel.

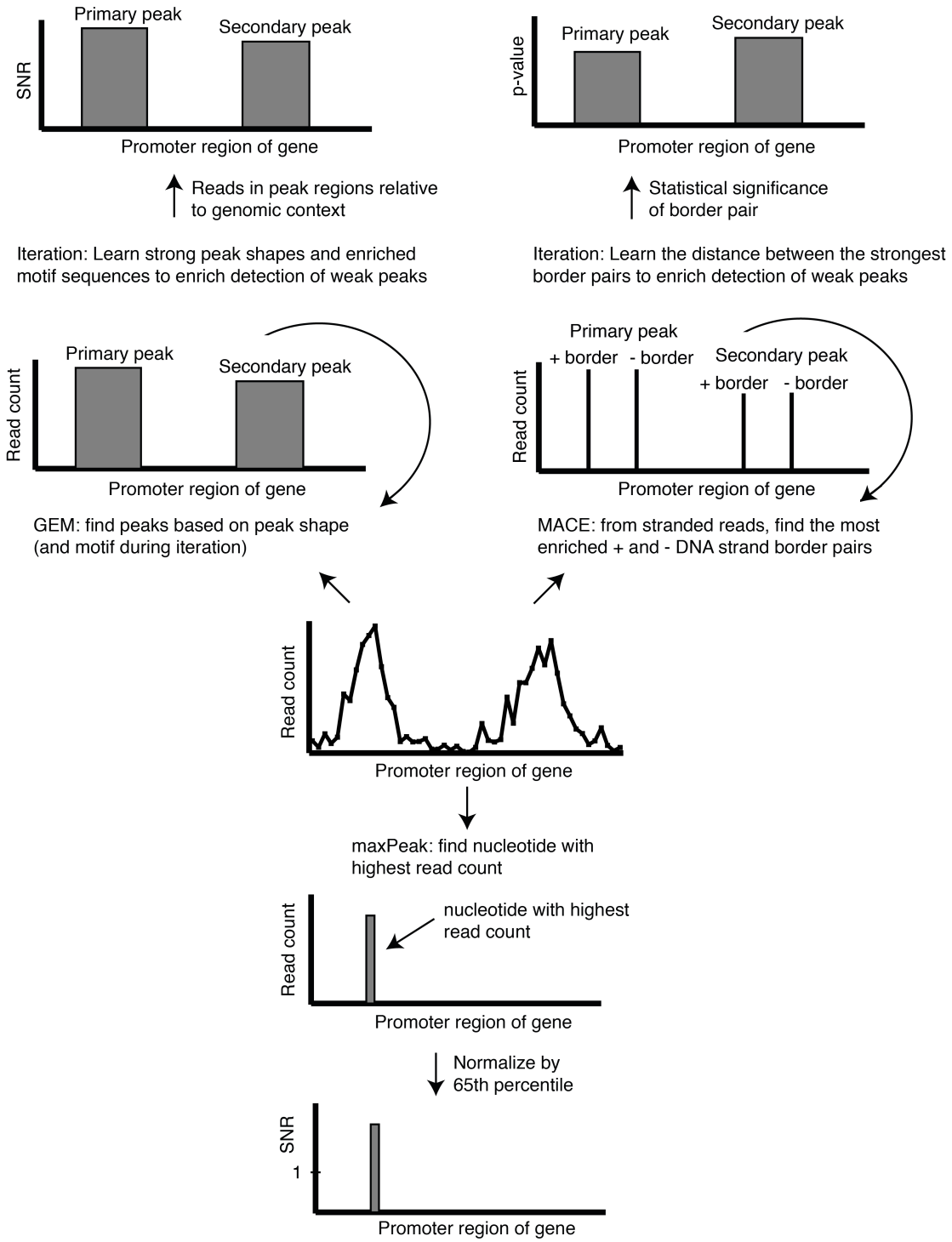


Figure S3.1: Illustration of the principles behind the three peak detection methods (PDMs) employed in this work: GEM (top-left), MACE (top-right) and *maxPeak* (bottom). All methods start from sorted and indexed BAM files. The approaches taken by GEM [35] and MACE [36], are established. The *maxPeak* approach counts the number of reads on both DNA strands for each nucleotide and, subsequently, assigns the single nucleotide with the highest read count as the gene's signal. Finally, a signal-to-noise ratio (SNR) is calculated by normalizing the read count by the 65^{th} percentile of all genes with a read count > 0.

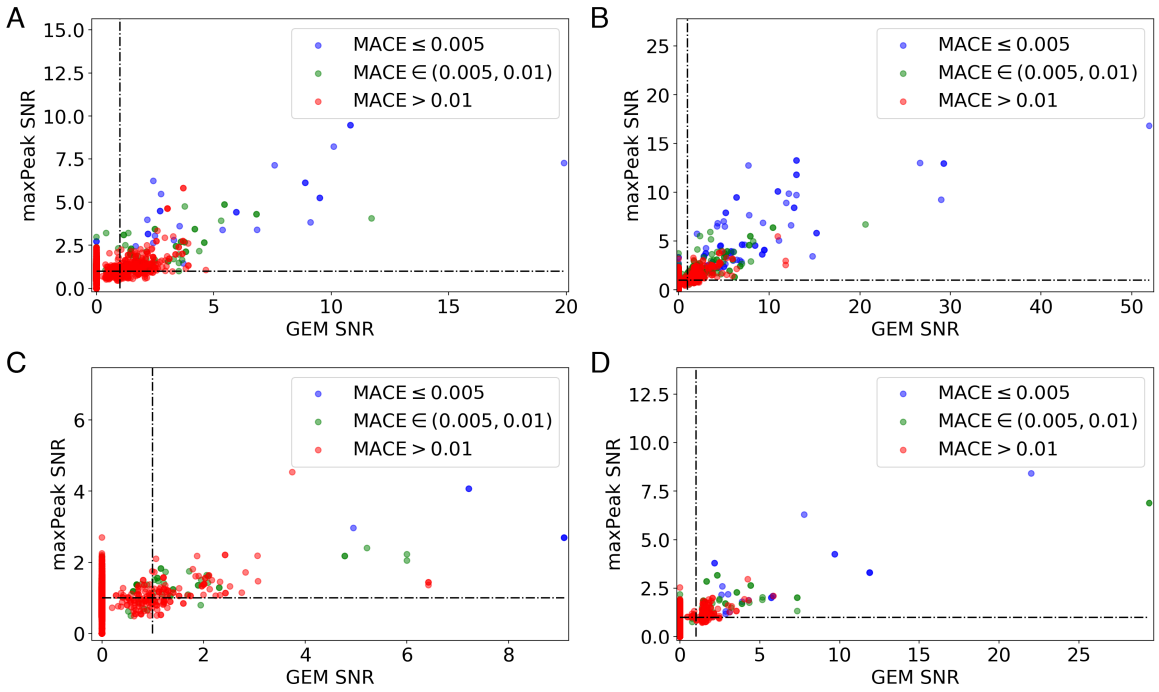


Figure S3.2: Plot of *maxPeak* SNRs vs. GEM SNRs for all genes with color indication for the MACE p-value. (**A**) Fkh1 exponential phase. (**B**) Fkh2 stationary phase. (**C**) Fkh2 exponential phase. (**D**) Fkh2 stationary phase. The horizontal and vertical black dotted lines represent the *maxPeak* and GEM target thresholds.

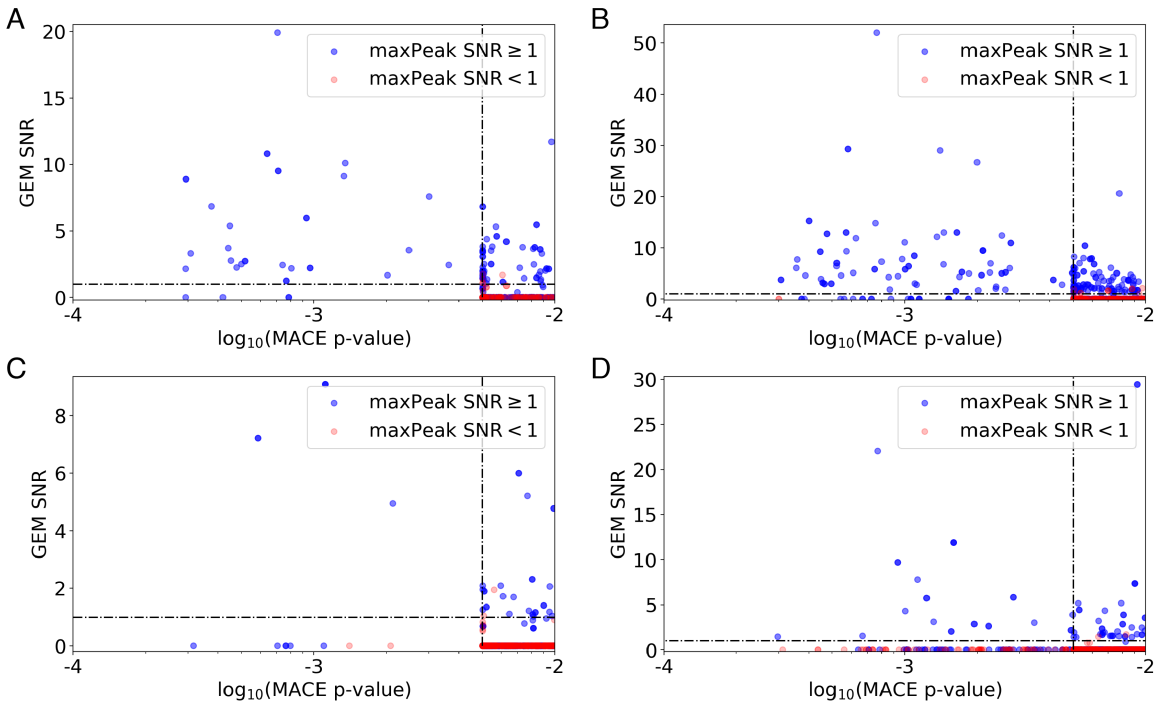


Figure S3.3: Plot of GEM SNRs vs. MACE SNRs for genes that scored a MACE p-value $<= 0.01$. (**A**) Fkh1 exponential phase. (**B**) Fkh2 stationary phase. (**C**) Fkh2 exponential phase. (**D**) Fkh2 stationary phase. The horizontal and vertical black dotted lines represent the GEM and MACE target thresholds, respectively.

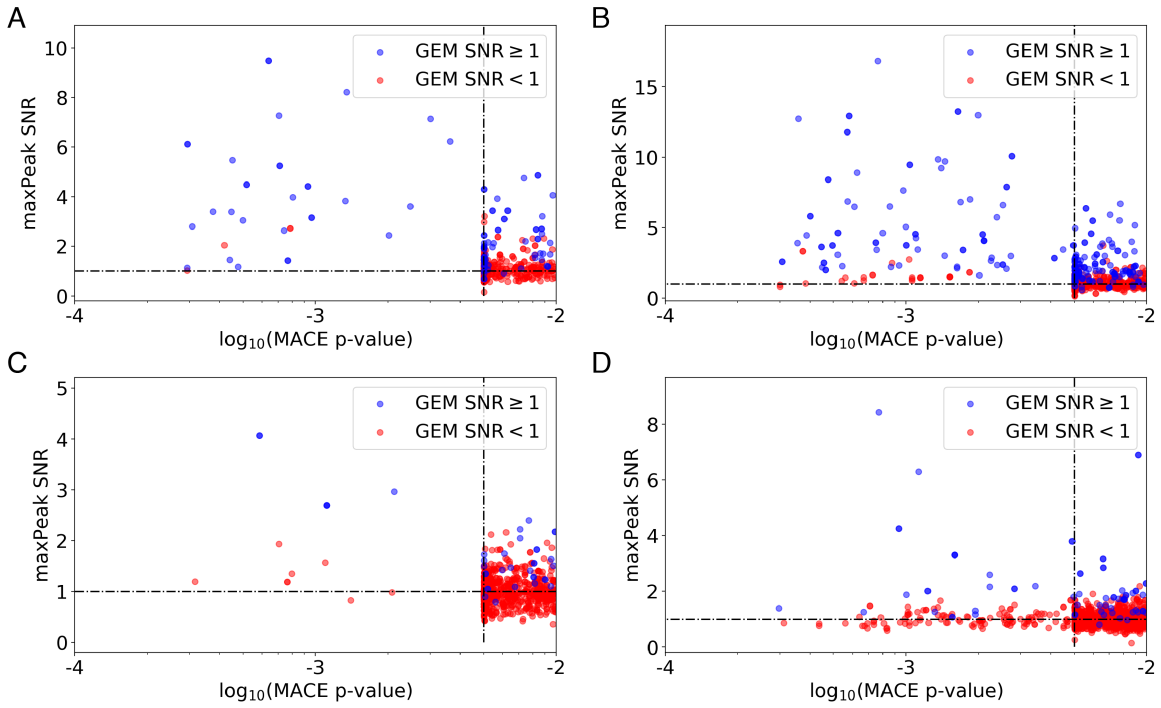


Figure S3.4: Plot of maxPeak SNRs vs. MACE p-values for genes that scored a MACE p-value <= 0.01. (**A**) Fkh1 exponential phase. (**B**) Fkh2 stationary phase. (**C**) Fkh2 exponential phase. (**D**) Fkh2 stationary phase. The horizontal and vertical black lines represent the maxPeak and MACE target thresholds, respectively.

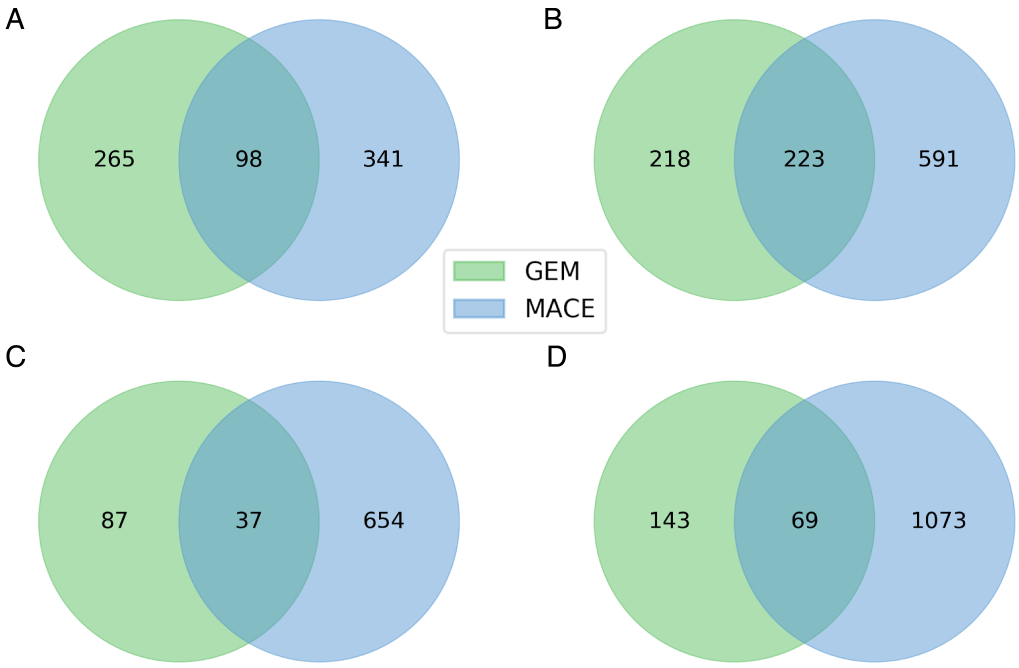


Figure S3.5: 2-way Venn diagram of target gene overlap among the two PDMs: GEM and MACE. For this image genes are considered targets when the SNR ≥ 1 (GEM) and the p-value ≤ 0.01 (MACE). (**A**) Fkh1 exponential phase. (**B**) Fkh1 stationary phase. (**C**) Fkh2 exponential phase. (**D**) Fkh2 stationary phase.

Supplementary Text

Divergence in gene targets retrieved by GEM vs. MACE

Fig. S3.5 highlights the divergence in the targets retrieved by GEM and MACE when using thresholds of 1 and 0.01, respectively. The majority of genes retrieved by either GEM or MACE are not recovered by the other method, although there is a common set of targets predicted by both PDMs. This divergence led us to develop the novel *maxPeak* method to decide whether to consider the non-overlapping genes shown in Fig. S3.5 as targets.

Comparing target genes retrieved by *maxPeak* vs. GEM vs. MACE

We explored the divergence in target genes retrieved between *maxPeak*, GEM and MACE. In Fig. S3.6 the Venn diagrams of the number of overlapping target genes are shown. Strikingly, there is a large number of genes retrieved by only one out of the three PDMs for all four experimental conditions. We emphasize that the divergence is not just between GEM and MACE, on the one hand, and *maxPeak*, on the other hand, but actually between all the three PDMs. This result may point to different features of the signal being retrieved by the three PDMs. Therefore, we explicitly required that target genes need to be retrieved by at least two out of three PDMs. Of note, by design, target genes retrieved by both GEM and MACE were also retrieved by *maxPeak* (see Supplementary Materials and Methods), and no overlap was observed between MACE and GEM.

Target genes identified by all the three methods

We observed that a small subset of target genes is in common among all the three peak detection methods, i.e. *maxPeak*, GEM and MACE. Thus, we refer to these genes as '3x PDM verified'. This subset consists of 31, 84, 6 and 25 genes for Fkh1 exponential, Fkh1 stationary, Fkh2 exponential and Fkh2 stationary, respectively. In total, these add up to 112 unique genes, 96 and 27 for Fkh1 and Fkh2, respectively. The '3x PDM verified' targets are particularly interesting as these are genes that are reported as significant targets regardless of the applied PDM. In Table S3.1 , 3x PDM verified target genes per experimental condition are listed with their standard name, if available.

Quantification of agreement between ChIP-exo peak locations and published ChIP-chip enriched regions

We downloaded the Supplementary Tables S2 and S4 provided by Ostrow et al. 2014, listing all significant binding events and the upstream regions of gene ORFs that they considered. We then filtered out the significant binding events that fall within the upstream region of gene ORFs, and plotted these for each chromosome

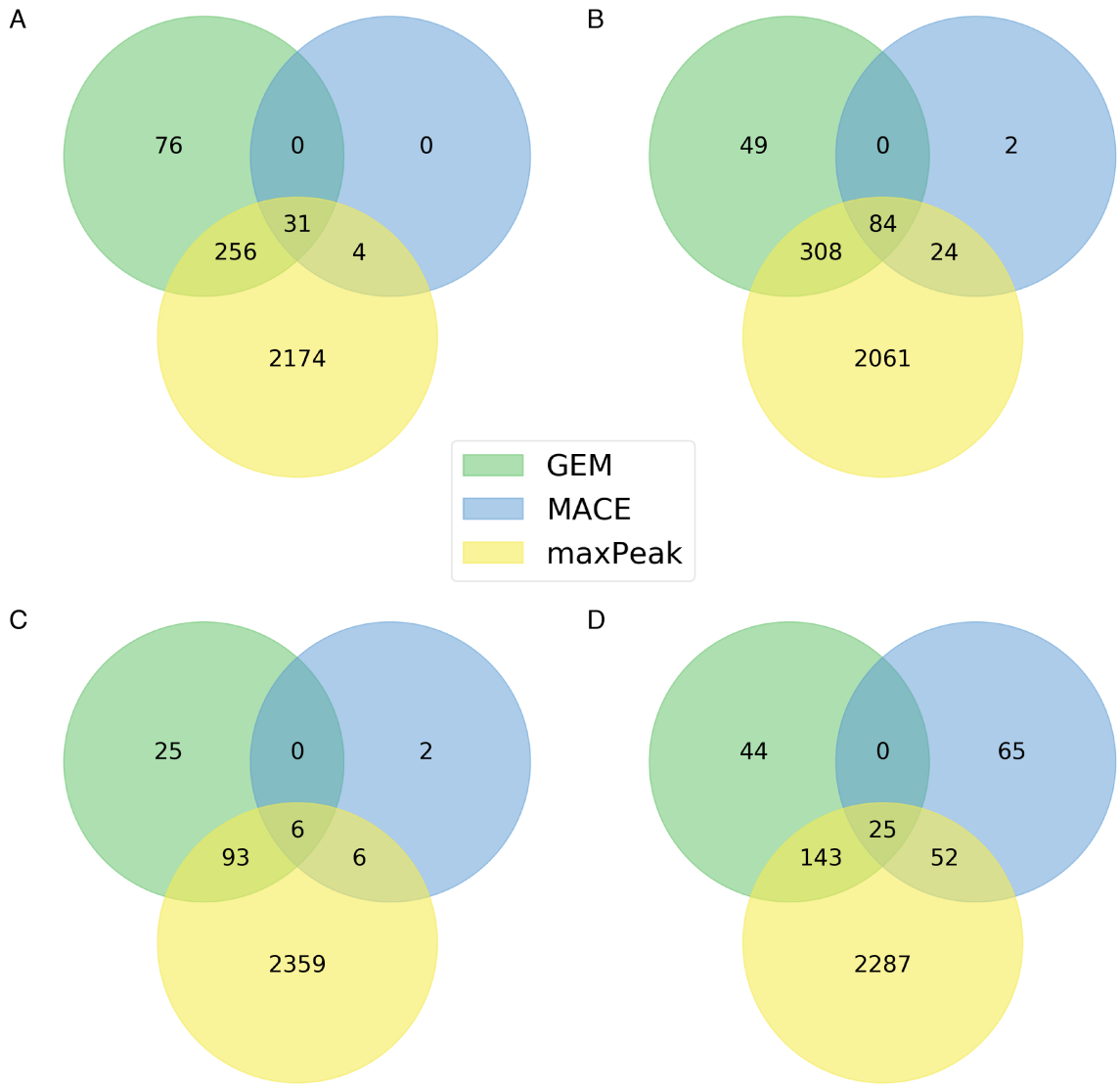


Figure S3.6: 3-way Venn diagrams of the overlap of the number of target genes retrieved among the three peak detection methods: *maxPeak*, GEM and MACE. In this image target genes are defined as having SNR ≥ 1 (*maxPeak*), the SNR ≥ 1 (GEM) and the p-value ≤ 0.005 (MACE). (**A**) Fkh1 exponential phase. (**B**) Fkh1 stationary phase. (**C**) Fkh2 exponential phase. (**D**) Fkh2 stationary phase.

and each Forkhead transcription factor separately. Finally, we overlapped those with the peak locations that we identified using maxPeak, GEM and MACE.

We summarized all the peak location comparisons in single figures for each chromosome and Forkhead transcription factor, and for both exponential and stationary phases. Fig. 3.3 in the main text shows the summary plots for Fkh2 in exponential phase for chromosome XVI, which contains CLB2, a major Fkh2 target. We note that Ostrow et al. 2014 performed the ChIP-chip experiments in exponential phase; however, we believe that images relative to the stationary phase may provide additional insight (see Supplementary exo-chip Figures for the complete set of summary plots).

Experiment	Target genes retrieved by *maxPeak*, GEM and MACE
Fkh1 exponential	**ALG5**, BRN1, CLN1, DIN7, ECM10, EGO2, **ERS1**, **EXG1**, FHL1, FIN1, FIR1, FLR1, FRK1, **HOS3**, **IDI1**, KIP2, NEW1, OSH7, PES4, RCR2, **RNR1**, RPS1B, **SAS3**, **SUN4**, TEM1, **UBP16**, VIK1, YDR003W-A, **YKL069W**, YLR299C-A, YPL251W
Fkh1 stationary	**ABP140**, **ADH4**, AIM46, **ALG5**, **APA1**, **ARK1**, ARP7, BFA1, BRN1, **BUB1**, BUD3, BUD8, **CHA1**, CIK1, **CIT2**, CLB4, **CPR7**, CSI1, CSN9, DCC1, **DIT1**, **DIT2**, DSE2, ECM10, **ERG26**, FHL1, FIN1, FIR1, FLR1, FRK1, GAS3, GEA1, **HOS3**, **HTS1**, **HXT5**, **IDI1**, **MKK2**, **MNT2**, MUM3, NBL1, NDD1, NEW1, NUD1, NUP2, NVJ3, OSH7, PDS5, PES4, **PMA1**, **PNP1**, **RNR1**, RPL39, RPL8A, RPN10, RPN11, RPS1B, RPS22A, **SAS3**, SCW11, SFG1, SGO1, SPC24, TDA7, TEM1, TIF1, TIM18, TMN2, **TOP2**, TRS85, VAC17, VIK1, VTI1, WTM1, YCS4, YGL007C-A, YGR111W, YLR334C, YMC2, YNL089C, YNL174W, YOR314W-A, YPI1, YPL251W, YPT31
Fkh2 exponential	CLB2, ENV9, **PMA1**, SRL1, YGL007C-A, YOR248W
Fkh2 stationary	ASE1, BUD4, **CHS2**, CLB2, CLN1, ECM10, EIS1, ENV9, FDO1, FRK1, **HOS3**, **IDI1**, JSN1, KIP2, MTC6, OSH7, RGI2, SCO1, SFG1, SPO12, SRL1, VHR1, YBR138C, YOR248W, YOR314W-A

Table S3.1: 3x PDM verified target genes, i.e. target genes retrieved by all three peak detection methods discussed in this work. Genes encoding an enzyme are indicated in bold. Genes that have not been reported as Fkh targets in previous ChIP studies [25] and/or [24] are indicated in red color.

We quantified the overlap in peak events upstream of gene ORFs between our study and Ostrow et al. 2014. Since in the latter study a window up to 500 bp upstream of the start of an ORF is considered, as opposed to 1000 bp windows considered in our work, we counted the number of genes with peaks within any enriched region identified by Ostrow et al. 2014. However, we observed no additional overlap for Fkh2 and only four additional genes that overlapped for Fkh1. As shown in Table S3.2 , between 52% – 82% of all our target genes fall into enriched regions as identified by Ostrow et al. 2014. The 18% – 48% of genes that show a peak outside of enriched regions identified by Ostrow et al. 2014 represent target genes that we identified, but that they did not. These results highlight the increased specificity achieved using ChIP-exo as compared to ChIP-chip. Vice versa, 43% – 51% of the enriched regions identified by Ostrow et al. 2014 that were upstream of an ORF contained at least one significant ChIP-exo peak (see Table S3.3). This percentage reflects the higher specificity of ChIP-exo and the higher stringency applied in this work.

	Fkh1 exponential	Fkh1 stationary	Fkh2 exponential	Fkh2 stationary
No overlap	54	76	43	106
Overlap	237	340	62	114
Total	291	416	105	220
% Overlap	81 %	82 %	59 %	52 %

Table S3.2: Quantification of the number of peak events for target genes identified in this work, which do or do not overlap with enriched regions identified by Ostrow et al. 2014.

	Fkh1 exponential	Fkh1 stationary	Fkh2 exponential	Fkh2 stationary
% ChIP-chip regions containing a ChIP-exo peak	51 %	50 %	46 %	43 %

Table S3.3: Quantification of the number of ChIP-chip enriched regions identified by Ostrow et al. 2014 that contain at least one significant ChIP-exo peak.

Functional category	Genome-wide %	Fkh1 exponential %	Fkh1 stationary %	Fkh2 exponential %	Fkh2 stationary %
None	49.58	-5.93	-6.31	-8.63	-9.12
Metabolism	41.29	-2.46	-2.35	-10.82	-0.84
Cell cycle	4.89	6.45	6.41	12.25	7.38
Signal transduction	2.29	0.81	0.84	0.57	-0.01
Cell division	1.19	1.56	0.97	7.38	2.44
DNA replication	0.76	-0.42	0.44	0	0.15

Table S3.4: Global function enrichment of Fkh target genes compared to the set of all ORFs. Percentages in the last four columns are calculated with respect to the "Genome-wide %" column.

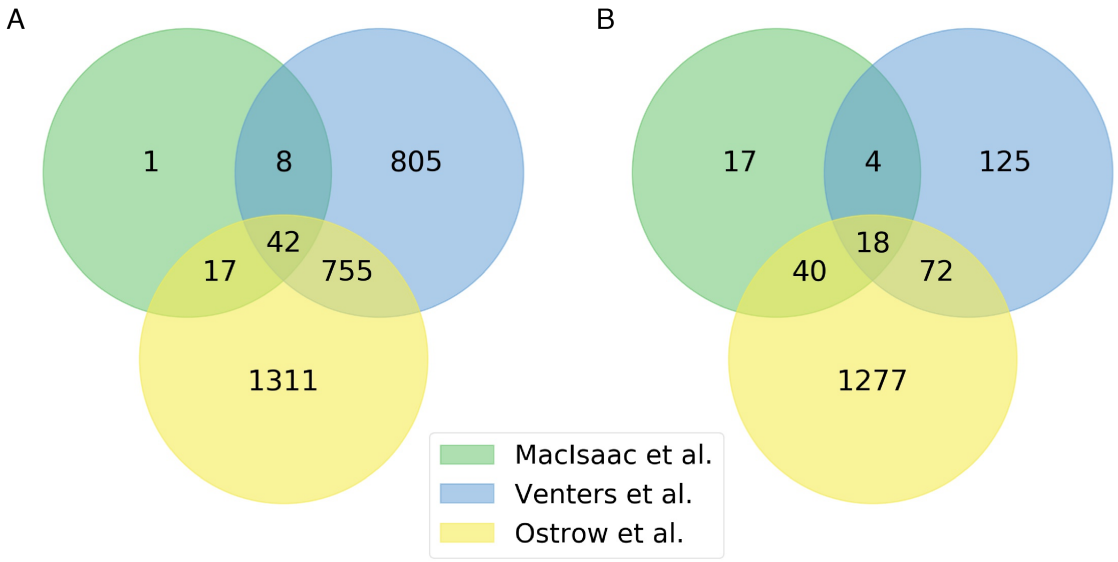


Figure S3.7: 3-way Venn diagrams of the overlap of the number of target genes retrieved among available ChIP-chip studies. (**A**) Fkh1 exponential phase. (**B**) Fkh2 exponential phase.

Overlap of target genes predicted by MacIsaac *et al.*, 2006, Venters *et al.*, 2011 and Ostrow *et al.*, 2014

In Fig. S3.7 the overlap among retrieved target genes between the three available ChIP-chip studies [24–26] is shown.

Cell cycle-regulated target genes identified in stationary phase

In Fig. S3.8, cell cycle-regulated target genes of Fkh1 and Fkh2 in the stationary phase ChIP-exo experiments are shown as a stack plot (analogously to the cell cycle-regulated target genes in exponential phase shown in Fig. 3.5).

CDC28 as an informative example for expected expression windows of Fkh1 and Fkh2 target genes

If the expression of target genes that we identified is actually regulated by Fkh1 and Fkh2, those genes would more likely exhibit expression peaks in a window surrounding, but most likely immediately after, the expression peaks of Fkh1 and Fkh2. However, this window of gene expression is subjected to noise, since RNA levels may peak after a delay that may be due to several reasons: (i) the transcription factor(s) may not be the only ones regulating the promoter of target genes; (ii) the transcription factor(s) may function in a transcriptional complex where other components bind at a different timing; (iii) the transcription factor(s) may be secondary activating elements that cause only subtle changes in the expression level of target genes; and (iv) the timing of the peak of expression depends on the mRNA half-life.

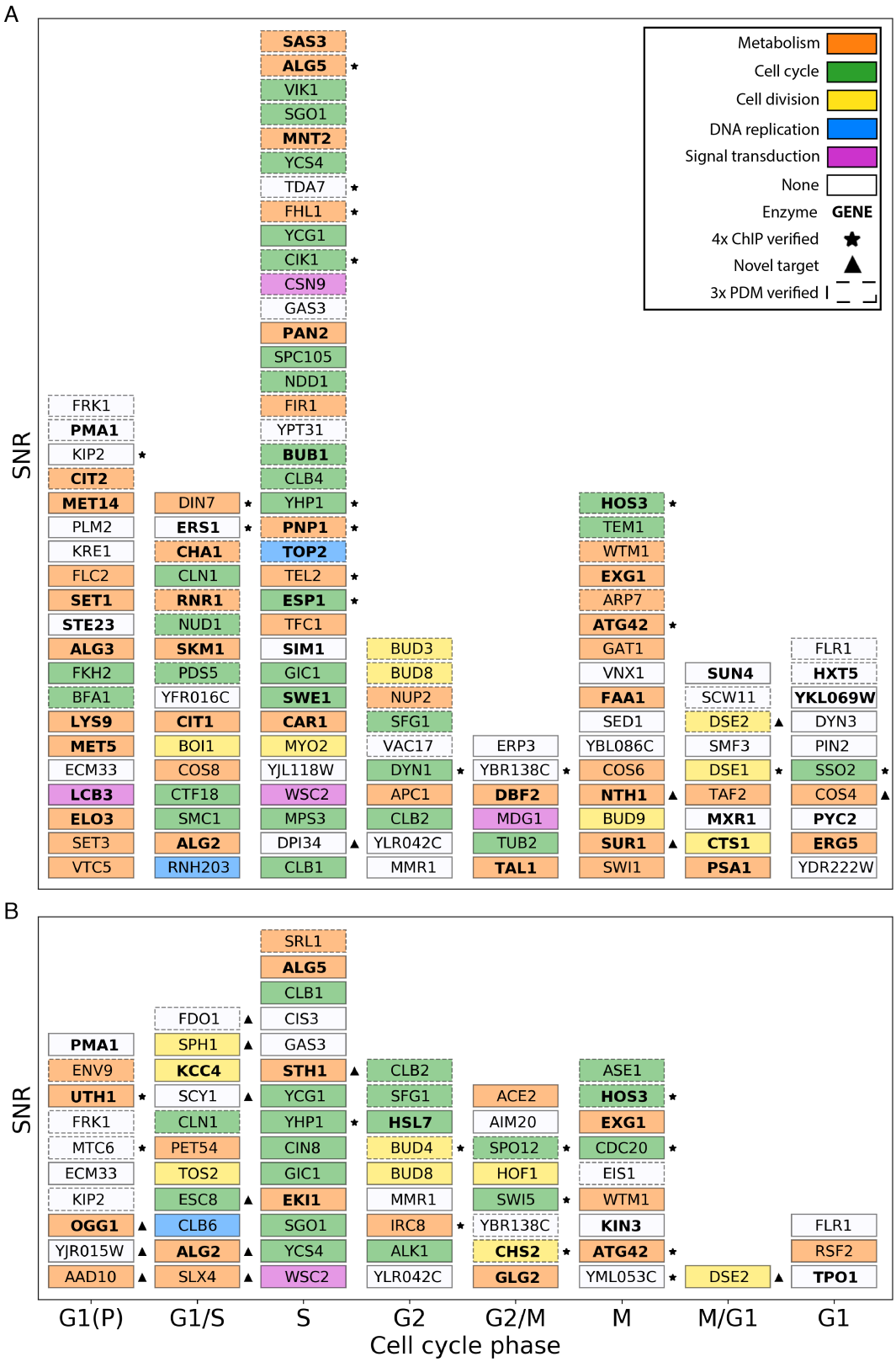


Figure S3.8: Stack plot of targets identified by ChIP-exo in stationary phase that have a cell cycle-regulated peak expression level [18]. (**A**) Fkh1 target genes. (**B**) Fkh2 target genes. Within each column a higher position on the y-axis indicates

Figure S3.8: (Continued) a higher *maxPeak* SNR. The x-axis indicates the phases of peak expression of genes, as reported [18]. The color for each gene indicates its major biological function if identified in GEMMER [31]. Targets marked with an asterisk are verified by all four ChIP studies, whereas targets marked with a triangle indicate novel target genes that have not been reported in the previous ChIP studies. Targets identified as significant by all three PDMs are shown with a dashed border.

Rowicka et al. [18] reported the timing difference between the CDC28 expression peak and the expression peak of Cdc28 targets see Fig. 4 in [18]. CDC28 showed two expression peaks at 29.5 and 154 minutes. Generally, when the expression of a transcription factor begins to increase but is not yet at its peak, activation of target genes may already initiate and then continue until after the peak has occurred and the transcription factor is degraded. The distribution of CDC28 target gene expression showed two peaks as well, roughly around 10 – 75 minutes and 125 – 175 minutes, respectively. Expression of target genes appears to initiate about 25 minutes prior the CDC28 expression peak is reached, and continues up to 45 minutes after the peak has occurred.

Global GO term analysis of Fkh target genes

In GEMMER the GO term annotations for all protein-coding genes listed in SGD (Saccharomyces Genome Database) are retrieved and traced back through the hierarchical tree of GO terms to one of the following high-level GO terms: *Cellular metabolic process* (GO:0044237), *Cell cycle* (GO:0007049), *Cell division* (GO:0051301), *Signal transduction* (GO:0007165), and *DNA replication* (GO:0006260). These terms fall under the GO term *Cellular process*, with *DNA replication* falling under *Cellular metabolic process*. Each such GO term annotation is assigned to one of the high-level terms listed above. For each gene, GEMMER adds up the number of annotations that fall under each high-level GO term. The GO term with the highest count is then assigned as the gene's primary function.

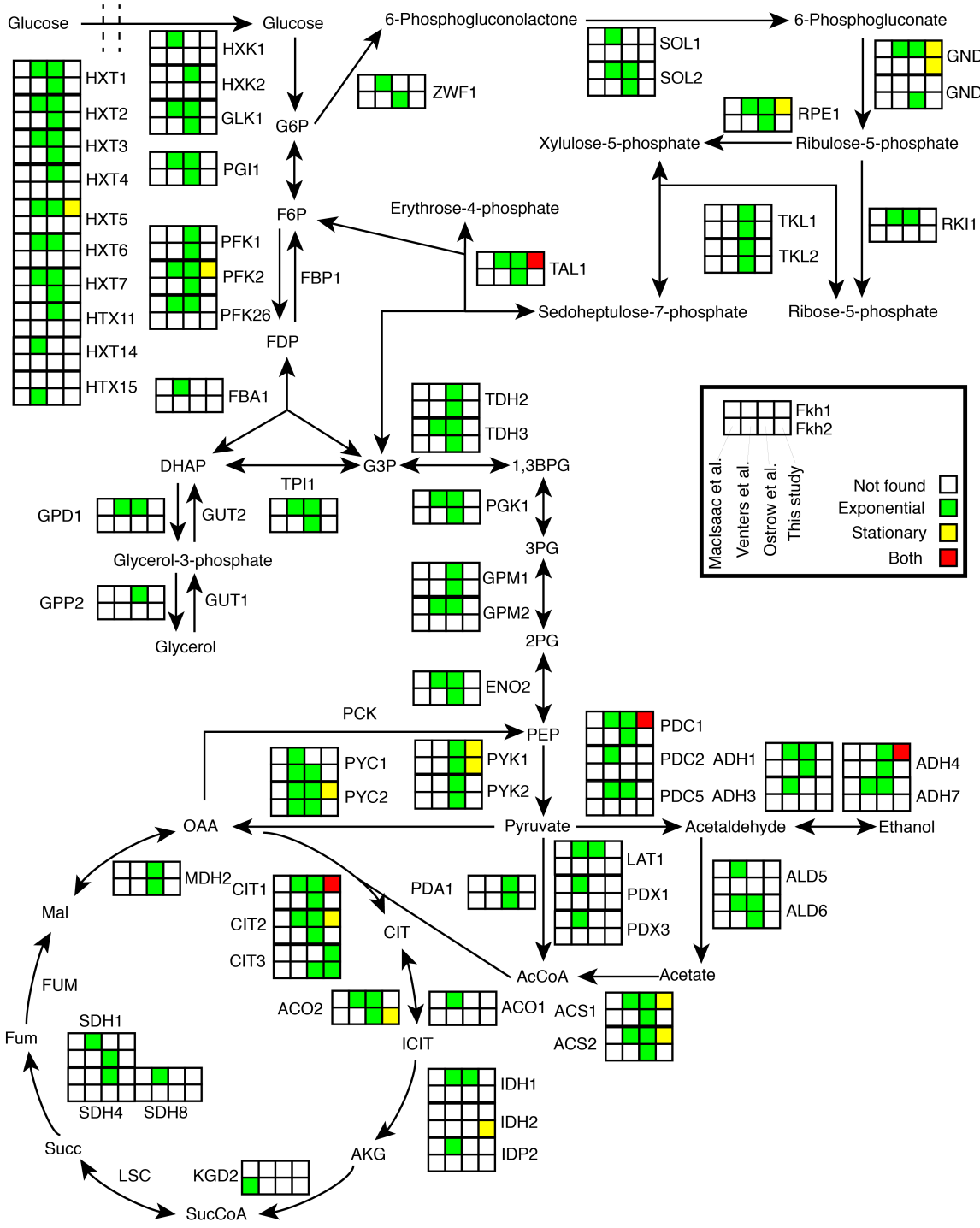


Figure S3.9: Overview of metabolic enzymes in central carbon metabolism that are targets of Fkh1 or Fkh2, as indicated by the four available ChIP studies. Each enzyme is associated with eight squares divided in two rows (Fkh1, top row; Fkh2, bottom row) representing data analysis of four different genome-wide studies: [24–26] and this study. Empty squares indicate that a gene was not retrieved as a significant target, whereas colored squares indicate positive evidence. Differences between the results in exponential and stationary phases are visualized through the color of the squares (see the figure insert). Isoenzymes without

Figure S3.9: (Continued) binding evidence in any of the three available ChIP studies were neglected. Some metabolic enzymes have no associated squares due to absence of an isoenzyme with an experimental validation.

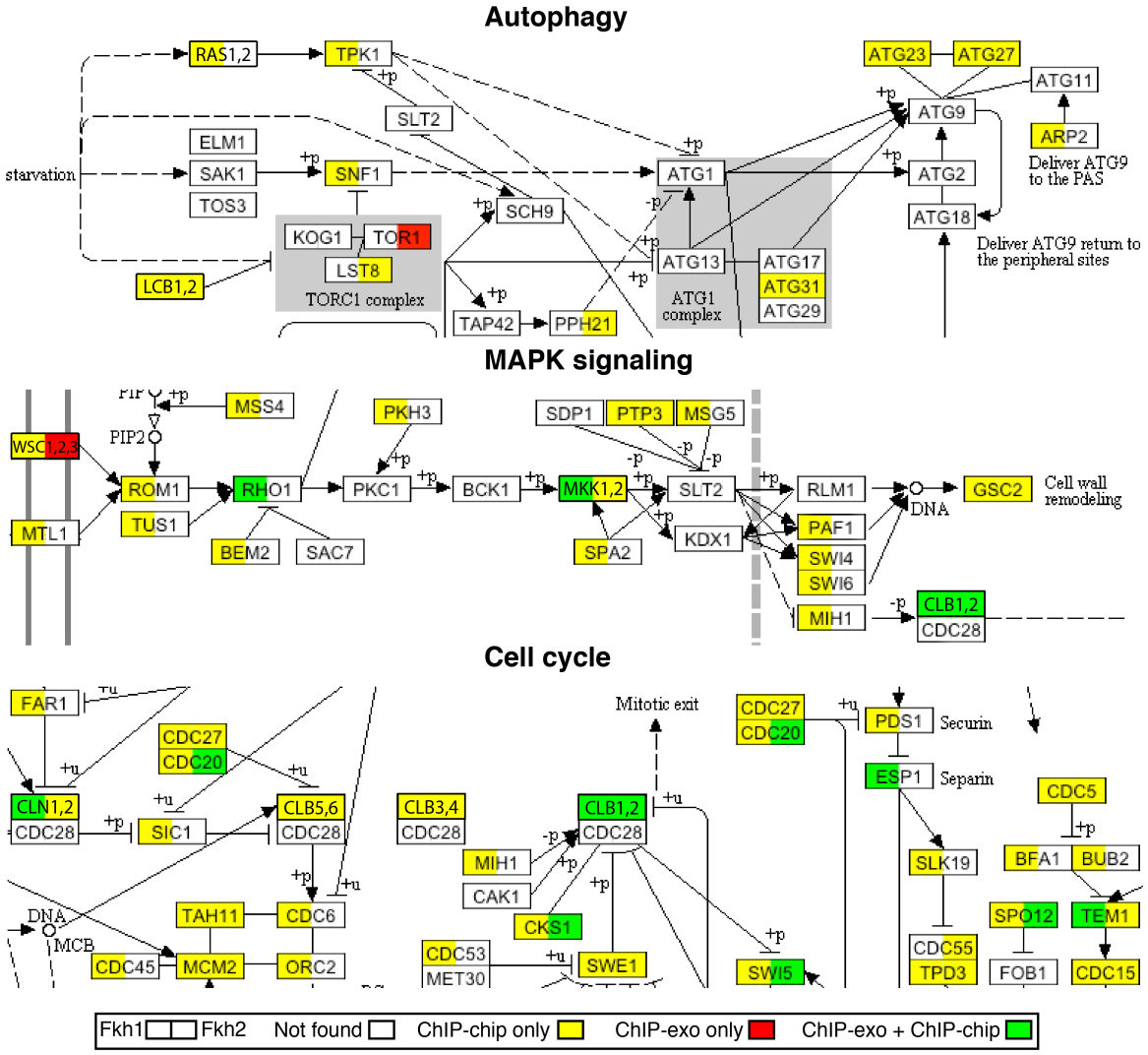


Figure S3.10: Fkh1 and Fkh2 target genes with wide-ranging functionality shown in KEGG pathways. Here visualized are KEGG maps of autophagy, MAPK signaling and cell cycle (see Supplementary KEGG Figures for all 25 pathways). The left half of the rectangle representing each gene is associated with Fkh1 results, whereas the right half is associated with Fkh2 results. Colors indicate available experimental evidence: targets shown in one of the three ChIP-chip studies (yellow); targets shown only by the ChIP-exo study (red); and targets shown by both ChIP-chip and ChIP-exo studies (green). The color of boxes representing two genes reflects the highest score retrieved for those genes.

In addition to the formal enrichment test of GO terms among the targets (see main text), the genes identified in our ChIP-exo experiments were analyzed for their functional enrichment, focusing on the global rather than on the specific

functions (see Table S3.4). From the analysis we observed that Fkh1 and Fkh2 show an increased percentage of target genes that fall within the GO terms *Cell cycle* and *Cell division*, both in exponential and in stationary phases, when compared with the set of all protein-coding genes. In all experiments except Fkh2 stationary phase the percentage of genes with a function in *Signal Transduction* was also increased. In contrast, in all four experiments the percentage of target genes with a function in metabolism was lower than the genome-wide percentage. In all, the GO term analyses support the earlier finding that Fkh targets are primarily cell cycle genes [8].

Summary of metabolic functioning of Fkh target genes

We highlighted the target genes that are metabolic enzymes in central carbon metabolism. Fig. S3.9 explicitly shows Fkh target genes retrieved by the previous ChIP studies [24–26] as well as by the ChIP-exo datasets generated in this study.

Functionality of Fkh target genes across KEGG pathways

Fig. S3.10 highlights the wide-reaching functions of the Fkh1 and Fkh2 target genes, visualizing examples of the Autophagy, MAPK signaling and Cell cycle pathways as reported in KEGG.

Distribution of Fkh targets across KEGG Pathways

The KEGG Pathways were the target genes of Fkh1 and Fkh2 occur were analyzed. In Table S3.5 , the genes occurring in a selected subset of the KEGG Pathways are reported. We selected a subset of the 89 pathways that highlighted the diversity of target functions across the cell cycle (cell cycle and meiosis), signaling (MAPK signaling and mitophagy), core metabolic pathways (glycolysis, TCA cycle, oxidative phosphorylation and biosynthesis of amino acids) and RNA and protein synthesis, transport and degradation (ribosome (biogenesis), proteasome, RNA degradation and RNA transport).

	Fkh1 exponential	Fkh1 stationary	Fkh2 exponential	Fkh2 stationary
Cell cycle - yeast	BRN1, **BUB1**, CLB1, CLN1, **DBF2**, **ESP1**, SCC4, SMC1, TEM1, TUP1, YCG1, YCS4	APC1, BFA1, BRN1, **BUB1**, CDC26, **CDC7**, CLB1, CLB2, CLB4, CLN1, **DBF2**, **ESP1**, ORC6, PHO4, SMC1, **SWE1**, TEM1, YCG1, YCS4, YHP1	CDC20, CKS1, CLB1, CLB2, **HSL7**, MAD1, SPO12, SWI5, YCS4, YHP1	CDC20, CDC26, CKS1, CLB1, CLB2, CLB6, CLN1, **HSL7**, **KCC4**, ORC3, PHO4, SPO12, SWI5, YCG1, YCS4, YHP1
MAPK signaling pathway	CLB1, CLN1, **HOG1**, **MKK2**, RHO1, **RSP5**, TUP1	CLB1, CLB2, CLN1, **HOG1**, **MKK2**, RHO1, **RSP5**, **SKM1**, **SWE1**, WSC2	CLB1, CLB2, **HSL7**, SLG1	CLB1, CLB2, CLB6, CLN1, **HSL7**, WSC2
Mitophagy - yeast	**HOG1**, **MKK2**	**HOG1**, **MKK2**, MMM1, WSC2	SLG1, **TOR2**	WSC2
Glycolysis/Gluconeogenesis	**ADH4**, **IRC15**, **PDC1**	**ACS1**, **ACS2**, **ADH4**, **CDC19**, **IRC15**, **PDC1**, **PFK2**		**CDC19**
Citrate cycle (TCA cycle)	**CIT1**, CIT3, **IRC15**	**CIT1**, **CIT2**, **IRC15**, **PYC2**	**CIT3**	**IDH2**
Oxidative phosphorylation	**PMA1**	**ATP14**, **COX4**, **PMA1**	**PMA1**	**PMA1**
Biosynthesis amino acids	**ARG5,6**, **CHA1**, **CIT1**, CIT3, **TAL1**	**ALT2**, **CAR1**, **CDC19**, **CHA1**, **CIT1**, **CIT2**, **GLN1**, **LYS9**, **PFK2**, **PYC2**, **RPE1**, **SER3**, **TAL1**	**CIT3**	**ACO2**, **CDC19**, **IDH2**
Ribosome biogenesis	FCF1, UTP4	**POP3**, UTP4		**POP3**, UTP4
Ribosome	MRPL10, MRPL32, RPL20B, RPL37B, RPL40B, RPL8A, RPS19A, RPS1B	MRPL32, RPL20B, RPL26B, RPL37B, RPL39, RPL40B, RPL8A, RPS19A, RPS1B, RPS21B, RPS22A		MRP17, RPL16A, RPL24A, RPL30
Proteasome	RPN10, RPN11, RPN12	RPN10, RPN11, RPN12		**PRE8**
RNA degradation	SKI6	CDC39, EDC3, LSM8, **PAN2**, **PFK2**, SKI6		
RNA transport	MLP1, SEH1, **SUB2**, TIF1	MLP1, **POP3**, SEH1, **SUB2**, TIF1, TIF34	**HSL7**	**HSL7**, **POP3**, TIF3

Table S3.5: KEGG Pathway occurrence of Fkh target genes. Genes encoding an enzyme are indicated in bold. Genes that have not previously been observed as Fkh targets in previous ChIP studies [24–26] are indicated in red. The listed KEGG Pathways are not complete, but form a representative subset out of 89 pathways that contain one or more Fkh target genes.

Overlap in Fkh target genes identified in ChIP-chip studies

In Table S3.6 we report the overlap of Fkh target genes identified using ChIP-chip studies [24–26]. We observe that most of the Fkh2 targets are recovered as Fkh1 targets, although we note that the number of targets reported [24, 25] is significantly higher for Fkh1 than Fkh2.

	MacIsaac et al. 2007	Venters et al. 2011	Ostrow et al. 2014
Common targets	27	176	1081
Unique Fkh1 targets	41	1434	1044
Unique Fkh2 targets	52	43	326

Table S3.6: Common and unique targets of Fkh1 and Fkh2 in three published ChIP-chip studies.

References

[1] A.-L. Barabási and Z. N. Oltvai. "Network biology: understanding the cell's functional organization". *Nature Reviews Genetics* 5 (2004), pp. 101–113. 10.1038/nrg1272.

[2] H. Kitano. "Computational systems biology". *Nature* 420 (2002), pp. 206–210. 10.1038/nature01254.

[3] F. Castiglione *et al.* "Modeling biology spanning different scales: an open challenge." *BioMed research international* 2014 (2014), p. 902545. 10.1155/2014/902545.

[4] W. Link and P. J. Fernandez-Marcos. "FOXO transcription factors at the interface of metabolism and cancer". *International Journal of Cancer* 141 (2017), pp. 2379–2391. 10.1002/ijc.30840.

[5] G. Murtaza *et al.* "FOXO Transcriptional Factors and Long-Term Living". *Oxidative Medicine and Cellular Longevity* 2017 (2017), pp. 1–8. 10.1155/2017/3494289.

[6] M. L. Golson and K. H. Kaestner. "Fox transcription factors: from development to disease". *Development* 143 (2016), pp. 4558–4570. 10.1242/dev.112672.

[7] E. W. Lam, J. J. Brosens, A. R. Gomes, and C.-Y. Koo. "Forkhead box proteins: tuning forks for transcriptional harmony". *Nature Reviews Cancer* 13 (2013), pp. 482–495. 10.1038/nrc3539.

[8] P. T. Spellman *et al.* "Comprehensive Identification of Cell Cycle–regulated Genes of the Yeast Saccharomyces cerevisiae by Microarray Hybridization". *Molecular Biology of the Cell* 9 (1998). Ed. by G. R. Fink, pp. 3273–3297. 10.1091/mbc.9.12.3273.

[9] D. Reynolds *et al.* "Recruitment of Thr 319-phosphorylated Ndd1p to the FHA domain of Fkh2p requires Clb kinase activity: a mechanism for CLB cluster gene activation." *Genes & development* 17 (2003), pp. 1789–802. 10.1101/gad.1074103.

[10] A. Pic *et al.* "The forkhead protein Fkh2 is a component of the yeast cell cycle transcription factor SFF." *The EMBO journal* 19 (2000), pp. 3750–61. 10.1093/emboj/19.14.3750.

[11] R. Kumar *et al.* "Forkhead transcription factors, Fkh1p and Fkh2p, collaborate with Mcm1p to control transcription required for M-phase". *Current Biology* 10 (2000), pp. 896–906. 10.1016/S0960-9822(00)00618-7.

[12] M. Koranda, A. Schleiffer, L. Endler, and G. Ammerer. "Forkhead-like transcription factors recruit Ndd1 to the chromatin of G2/M-specific promoters". *Nature* 406 (2000), pp. 94–98. 10.1038/35017589.

[13] J. A. Sherriff, N. A. Kent, and J. Mellor. "The Isw2 Chromatin-Remodeling ATPase Cooperates with the Fkh2 Transcription Factor To Repress Transcription of the B-Type Cyclin Gene CLB2". *Molecular and Cellular Biology* 27 (2007), pp. 2848–2860. 10.1128/MCB.01798-06.

[14] P. C. Hollenhorst *et al.* "Forkhead Genes in Transcriptional Silencing, Cell Morphology and the Cell Cycle: Overlapping and Distinct Functions for FKH1 and FKH2 in Saccharomyces cerevisiae". *Genetics* 154 (2000), pp. 1533–1548. 10.1093/genetics/154.4.1533.

[15] P. C. Hollenhorst, G. Pietz, and C. A. Fox. "Mechanisms controlling differential promoter-occupancy by the yeast forkhead proteins Fkh1p and Fkh2p: implications for regulating the cell cycle and differentiation". *Genes & Development* 15 (2001), pp. 2445–2456. 10.1101/gad.906201.

[16] C. Linke *et al.* "A Clb/Cdk1-mediated regulation of Fkh2 synchronizes CLB expression in the budding yeast cell cycle". *npj Systems Biology and Applications* 3 (2017), p. 7. 10.1038/s41540-017-0008-1.

[17] B.-J. Shi. "Decoding common and divergent cellular functions of the domains of forkhead transcription factors Fkh1 and Fkh2". *Biochemical Journal* 473 (2016), pp. 3855–3869. 10.1042/BCJ20160609.

[18] M. Rowicka, A. Kudlicki, B. P. Tu, and Z. Otwinowski. "High-resolution timing of cell cycle-regulated gene expression". *Proceedings of the National Academy of Sciences* 104 (2007), pp. 16892–16897. 10.1073/pnas.0706022104.

[19] T. I. Lee. "Transcriptional Regulatory Networks in Saccharomyces cerevisiae". *Science* 298 (2002), pp. 799–804. 10.1126/science.1075090.

[20] C. Linke *et al.* "Fkh1 and Fkh2 associate with Sir2 to control CLB2 transcription under normal and oxidative stress conditions". *Frontiers in Physiology* 4 (2013), pp. 1–17. 10.3389/fphys.2013.00173.

[21] S.-J. Lin *et al.* "Calorie restriction extends yeast life span by lowering the level of NADH." *Genes & development* 18 (2004), pp. 12–6. 10.1101/gad.1164804.

[22] Z. Hu, P. J. Killion, and V. R. Iyer. "Genetic reconstruction of a functional transcriptional regulatory network". *Nature Genetics* 39 (2007), pp. 683–687. 10.1038/ng2012.

[23] C. J. Viggiani, J. G. Aparicio, and O. M. Aparicio. "ChIP-Chip to Analyze the Binding of Replication Proteins to Chromatin Using Oligonucleotide DNA Microarrays". *DNA Replication*. Vol. 521. Humana Press, 2009, pp. 255–278. 10.1007/978-1-60327-815-7.

[24] A. Z. Ostrow *et al.* "Fkh1 and Fkh2 Bind Multiple Chromosomal Elements in the S. cerevisiae Genome with Distinct Specificities and Cell Cycle Dynamics". *PLoS ONE* 9 (2014). Ed. by Y. Wang, e87647. 10.1371/journal.pone.0087647.

[25] B. J. Venters *et al.* "A Comprehensive Genomic Binding Map of Gene and Chromatin Regulatory Proteins in Saccharomyces". *Molecular Cell* 41 (2011), pp. 480–492. 10.1016/j.molcel.2011.01.015.

[26] K. D. MacIsaac *et al.* "An improved map of conserved regulatory sites for Saccharomyces cerevisiae." *BMC bioinformatics* 7 (2006), p. 113. 10.1186/1471-2105-7-113.

[27] C. G. de Boer and T. R. Hughes. "YeTFaSCo: a database of evaluated yeast transcription factor sequence specificities". *Nucleic Acids Research* 40 (2012), pp. D169–D179. 10.1093/nar/gkr993.

[28] H. S. Rhee and B. F. Pugh. "ChIP-exo Method for Identifying Genomic Location of DNA-Binding Proteins with Near-Single-Nucleotide Accuracy". *Current Protocols in Molecular Biology*. Vol. 141. 4. Hoboken, NJ, USA: John Wiley & Sons, Inc., 2012, pp. 520–529. 10.1002/0471142727.mb2124s100.

[29] L. Ouyang *et al.* "Integrated analysis of the yeast NADPH-regulator Stb5 reveals distinct differences in NADPH requirements and regulation in different states of yeast metabolism". *FEMS Yeast Research* 18 (2018), pp. 1–12. 10.1093/femsyr/foy091.

[30] P. Holland *et al.* "Predictive models of eukaryotic transcriptional regulation reveals changes in transcription factor roles and promoter usage between metabolic conditions". *Nucleic Acids Research* 47 (2019), pp. 4986–5000. 10.1093/nar/gkz253.

[31] T. D. G. A. Mondeel, F. Crémazy, and M. Barberis. "GEMMER: GEnome-wide tool for Multi-scale Modeling data Extraction and Representation for Saccharomyces cerevisiae". *Bioinformatics* 34 (2018). Ed. by J. Wren, pp. 2147–2149. 10.1093/bioinformatics/bty052.

[32] G. Liu, D. Bergenholm, and J. Nielsen. "Genome-Wide Mapping of Binding Sites Reveals Multiple Biological Functions of the Transcription Factor Cst6p in Saccharomyces cerevisiae". *mBio* 7 (2016), e00559–16. 10.1128/mBio.00559-16.

[33] B. Langmead and S. L. Salzberg. "Fast gapped-read alignment with Bowtie 2". *Nature Methods* 9 (2012), pp. 357–359. 10.1038/nmeth.1923.

[34] H. Li *et al.* "The Sequence Alignment/Map format and SAMtools". *Bioinformatics* 25 (2009), pp. 2078–2079. 10.1093/bioinformatics/btp352.

[35] Y. Guo, S. Mahony, and D. K. Gifford. "High Resolution Genome Wide Binding Event Finding and Motif Discovery Reveals Transcription Factor Spatial Binding Constraints". *PLoS Computational Biology* 8 (2012). Ed. by S. Aerts, e1002638. 10.1371/journal.pcbi.1002638.

[36] L. Wang *et al.* "MACE: model based analysis of ChIP-exo". *Nucleic Acids Research* 42 (2014), e156–e156. 10.1093/nar/gku846.

[37] E. Kristiansson, M. Thorsen, M. J. Tamas, and O. Nerman. "Evolutionary Forces Act on Promoter Length: Identification of Enriched Cis-Regulatory Elements". *Molecular Biology and Evolution* 26 (2009), pp. 1299–1307. 10.1093/molbev/msp040.

[38] A. S. Kelley *et al.* "Determinants of Medical Expenditures in the Last 6 Months of Life". *Annals of Internal Medicine* 154 (2011), p. 235. 10.7326/0003-4819-154-4-201102150-00004.

[39] M. J. Herrgård *et al.* "A consensus yeast metabolic network reconstruction obtained from a community approach to systems biology". *Nature Biotechnology* 26 (2008), pp. 1155–1160. 10.1038/nbt1492.

[40] W. Luo and C. Brouwer. "Pathview: an R/Bioconductor package for pathway-based data integration and visualization". *Bioinformatics* 29 (2013), pp. 1830–1831. 10.1093/bioinformatics/btt285.

[41] C. T. Harbison *et al.* "Transcriptional regulatory code of a eukaryotic genome". *Nature* 431 (2004), pp. 99–104. 10.1038/nature02800.

[42] A. Kudlicki, M. Rowicka, and Z. Otwinowski. "SCEPTRANS: an online tool for analyzing periodic transcription in yeast". *Bioinformatics* 23 (2007), pp. 1559–1561. 10.1093/bioinformatics/btm126.

[43] R. J. Cho *et al.* "A Genome-Wide Transcriptional Analysis of the Mitotic Cell Cycle". *Molecular Cell* 2 (1998), pp. 65–73. 10.1016/S1097-2765(00)80114-8.

[44] R. R. Klevecz, J. Bolen, G. Forrest, and D. B. Murray. "A genomewide oscillation in transcription gates DNA replication and cell cycle". *Proceedings of the National Academy of Sciences* 101 (2004), pp. 1200–1205. 10.1073/pnas.0306490101.

[45] B. P. Tu. "Logic of the Yeast Metabolic Cycle: Temporal Compartmentalization of Cellular Processes". *Science* 310 (2005), pp. 1152–1158. 10.1126/science.1120499.

[46] T. Pramila *et al.* "The Forkhead transcription factor Hcm1 regulates chromosome segregation genes and fills the S-phase gap in the transcriptional circuitry of the cell cycle." *Genes & development* 20 (2006), pp. 2266–78. 10.1101/gad.1450606.

[47] H. Mi *et al.* "PANTHER version 11: expanded annotation data from Gene Ontology and Reactome pathways, and data analysis tool enhancements". *Nucleic Acids Research* 45 (2017), pp. D183–D189. 10.1093/nar/gkw1138.

[48] A. Z. Ostrow *et al.* "Conserved forkhead dimerization motif controls DNA replication timing and spatial organization of chromosomes in S. cerevisiae". *Proceedings of the National Academy of Sciences* 114 (2017), E2411–E2419. 10.1073/pnas.1612422114.

[49] T. L. Bailey *et al.* "MEME SUITE: tools for motif discovery and searching". *Nucleic Acids Research* 37 (2009), W202–W208. 10.1093/nar/gkp335.

[50] T. L. Bailey, N. Williams, C. Misleh, and W. W. Li. "MEME: discovering and analyzing DNA and protein sequence motifs". *Nucleic Acids Research* 34 (2006), W369–W373. 10.1093/nar/gkl198.

[51] T. L. Bailey. "DREME: motif discovery in transcription factor ChIP-seq data". *Bioinformatics* 27 (2011), pp. 1653–1659. 10.1093/bioinformatics/btr261.

[52] T. L. Bailey and P. Machanick. "Inferring direct DNA binding from ChIP-seq". *Nucleic Acids Research* 40 (2012), e128–e128. 10.1093/nar/gks433.

[53] S. Bandyopadhyay *et al.* "Rewiring of Genetic Networks in Response to DNA Damage". *Science* 330 (2010), pp. 1385–1389. 10.1126/science.1195618.

[54] E. Kuzmin *et al.* "Systematic analysis of complex genetic interactions". *Science* 360 (2018), eaao1729. 10.1126/science.aao1729.

[55] J. Sajman *et al.* "Degradation of Ndd1 by APC/CCdh1 generates a feed forward loop that times mitotic protein accumulation". *Nature Communications* 6 (2015), p. 7075. 10.1038/ncomms8075.

[56] G. Tuteja and K. H. Kaestner. "SnapShot:Forkhead Transcription Factors I". *Cell* 130 (2007), 1160.e1–1160.e2. 10.1016/j.cell.2007.09.005.

[57] G. Tuteja and K. H. Kaestner. "SnapShot:Forkhead Transcription Factors II". *Cell* 130 (2007), 1160.e1–1160.e2. 10.1016/j.cell.2007.09.005.

[58] H. Murakami, H. Aiba, M. Nakanishi, and Y. Murakami-Tonami. "Regulation of yeast forkhead transcription factors and FoxM1 by cyclin-dependent and polo-like kinases". *Cell Cycle* 9 (2010), pp. 3253–3262. 10.4161/cc.9.16.12599.

[59] W. Korver, J. Roose, and H. Clevers. "The winged-helix transcription factor Trident is expressed in cycling cells." *Nucleic acids research* 25 (1997), pp. 1715–9. 10.1093/nar/25.9.1715.

[60] J. Laoukili *et al.* "FoxM1 is required for execution of the mitotic programme and chromosome stability". *Nature Cell Biology* 7 (2005), pp. 126–136. 10.1038/ncb1217.

[61] J. Laoukili, M. Stahl, and R. H. Medema. "FoxM1: At the crossroads of ageing and cancer". *Biochimica et Biophysica Acta - Reviews on Cancer* 1775 (2007), pp. 92–102. 10.1016/j.bbcan.2006.08.006.

[62] J. Bloom and F. R. Cross. "Multiple levels of cyclin specificity in cell-cycle control". *Nature Reviews Molecular Cell Biology* 8 (2007), pp. 149–160. 10.1038/nrm2105.

[63] L. L. Breeden. "Cyclin transcription: Timing is everything". *Current Biology* 10 (2000), R586–R588. 10.1016/S0960-9822(00)00634-5.

[64] G. Zhu *et al.* "Two yeast forkhead genes regulate the cell cycle and pseudohyphal growth". *Nature* 406 (2000), pp. 90–94. 10.1038/35017581.

[65] J. M. Cherry *et al.* "Saccharomyces Genome Database: the genomics resource of budding yeast". *Nucleic Acids Research* 40 (2012), pp. D700–D705. 10.1093/nar/gkr1029.

[66] J. L. Y. Koh *et al.* "CYCLoPs: A Comprehensive Database Constructed from Automated Analysis of Protein Abundance and Subcellular Localization Patterns in Saccharomyces cerevisiae." *G3 (Bethesda, Md.)* 5 (2015), pp. 1223–32. 10.1534/g3.115.017830.

[67] W.-k. Huh *et al.* "Global analysis of protein localization in budding yeast". *Nature* 425 (2003), pp. 686–691. 10.1038/nature02026.

[68] S. Ghaemmaghami *et al.* "Global analysis of protein expression in yeast". *Nature* 425 (2003), pp. 737–741. 10.1038/nature02046.

CHAPTER 4

GEMMER: GEnome-wide tool for Multi-scale Modeling data Extraction and Representation for *Saccharomyces cerevisiae*

4.1 Introduction 132
4.2 Features 133
4.3 GEMMER's methodology 133
4.4 Conclusions 136

Adapted from:
T.D.G.A. Mondeel, F. Crémazy, M. Barberis, GEMMER: GEnome-wide tool for Multi-scale Modeling data Extraction and Representation for *Saccharomyces cerevisiae*, Bioinformatics. 34 (2018) 2147–2149. 10.1093/bioinformatics/bty052

"Data! Data! Data! I can't make bricks without clay!"

— Sir Arthur Conan Doyle [1]

Abstract

Multi-scale modeling of biological systems requires integration of various bits and pieces of information about genes and proteins that are connected together in networks. Spatial, temporal and functional information is available; however, it is still a challenge to retrieve and explore this knowledge in an integrated, quick and user-friendly manner. We present GEMMER (GEnome-wide tool for Multi-scale Modelling data Extraction and Representation), a web-based data-integration tool that facilitates high quality visualization of physical, regulatory and genetic interactions between proteins/genes in *Saccharomyces cerevisiae*. GEMMER creates network visualizations that integrate information on function, temporal expression, localization and abundance from various existing databases. GEMMER supports modeling efforts by effortlessly gathering this information and providing convenient export options for images and their underlying data.

4.1 Introduction

Biological systems are complex systems: they exist in space and time, and their behavior results from the coherent integration of functionally diverse elements that interact selectively and non-linearly [2]. The understanding has emerged that a cross-talk between molecular pathways is crucial to achieve the system's functions. In this context, generation of multi-scale models of biological systems, spanning multiple spatial, temporal and functional scales, is currently a major challenge in Systems Biology [3].

Crucial steps in multi-scale modeling are the identification and visualization of the biological function, and spatial localization of interactions that occur among a set of molecules. Tools that retrieve and visualize such interaction networks for several organisms exist. However, these are not specific for the budding yeast *Saccharomyces cerevisiae*, and do not combine the features of: (i) being web-based instead of a desktop application, (ii) allowing visual exploration through simultaneous clustering, colouring and filtering of molecules and their interactions that are (iii) based on function, localization, abundance, and timing at which they occur.

Here, we present GEMMER, a novel web-based data-integration and visualization tool for budding yeast that satisfies these three requirements. The tool provides unique features as compared to existing web-based visualization tools and databases (see Table 4.1 for a detailed comparison). Furthermore, through its export options, GEMMER conveniently integrates with external tools that may be used to build and simulate multi-scale models.

4.2 Features

GEMMER integrates (i) protein-coding genes, interactions and general and functional annotation from the Saccharomyces Genome Database (SGD) [4], (ii) localization and abundance data from both the CYCLoPs [5] and Yeast GFP Fusion Localization, YeastGFP [6, 7] databases, and (iii) the timing and cell cycle phase of peak occurrence of mRNA levels [8, 9]. GEMMER provides distinct webpages for each protein-coding gene, where this information may be viewed.

Features and information flow of GEMMER are summarized in Fig. 4.1. After the user selects one or more genes, for which the aforementioned information is retrieved, GEMMER generates an interaction network, which varies across functional, spatial and temporal scales. Nodes in this interaction network may be clustered and coloured based on their localization in a number of cellular compartments and their functional classification. In addition, interactions may be filtered out based on: (i) type of interaction (physical, genetic or regulation), (ii) total number of experiments suggesting them, (iii) unique experimental methodology, (iv) type of experimental evidence and (v) number of publications showing it. Similarly, nodes may be filtered out based on function (process or GO term), cellular compartment, and cell cycle phase where the peak of transcription occurs. As a result, the user receives as output an interaction network which is generated by using up-to-date literature data and filtered for their specific needs.

GEMMER provides a set of unique features as compared to existing web-based tools that allow visualization of budding yeast-specific data, i.e. STRING [10], BIOGRID [11], APID [12] and IntACT [13] (see Table 4.1 for a detailed comparison). These are: (i) Generating interaction networks seeded by > 1 protein; (ii) Filtering interactions on the number of unique experimental methods that have been employed to prove it; (iii) Clustering and colouring interactions based on cellular compartments or GO terms; (iv) Displaying protein expression levels; (v) Filtering nodes based on the network characteristics: degree, eigenvector or Katz centrality. Conversely, GEMMER currently lacks different visual layouts and certain export formats, such as PNG, JPEG and XML, features that are instead available in some of the aforementioned tools.

4.3 GEMMER's methodology

GEMMER stores the integrated data from the external databases in an SQLite database. The latter is updated by using a Python script that downloads data from the latest available releases of SGD, CYCLoPs, YeastGFP and SCEPTRANS databases. Periodic running of the update script provides GEMMER with up-to-date literature data.

The GEMMER front-end provides a user-friendly interface with a set of menus that facilitate user input. This includes the gene(s) of interest to build an interaction network, and the filtering, clustering, colouring and scaling of nodes in the visualized network. Upon querying, the input is processed by a PHP script that executes the core application. This has been written in Python

Feature	GEMMER	STRING	BioGRID	APID	IntACT
Interaction networks seeded by two or more nodes	X		X*		X
Browsing different organisms		X	X	X	X
Node size proportional to # interactions in the network	X		X		
Edge size proportional to # experiments used to show the interaction	X		X	X	
Coloring nodes based on their GO-term annotated function	X	X			
Clustering nodes based on their GO-term annotated function	X		X	X	
Clustering based on cellular compartments	X				
Different possible visual layouts		X	X**	X**	X**
Filtering on physical, regulatory and/or genetic interactions	X	X	X		
Filtering on number of experiments that show an interaction	X	X***		X	
Filtering on # unique experimental methods used to show interaction	X				
Filtering nodes: degree, eigenvector or Katz centrality	X				
Displaying peaks of transcription levels	X				
Link to original publication through PubMed	X	X	X		
Display PDB, Pfam, post-translational modifications		X	X		
Bitmap or vector image formats available for export	SVG	PNG, SVG	PNG	JPG, PNG	
Table formats available for export	XLSX	TSV, TXT, XML		TXT	

Table 4.1: Feature comparison between GEMMER and existing alternative platforms. * network visualization of user-supplied nodes only. ** visualizations via Cytoscape. *** confidence factor existing but not explicit

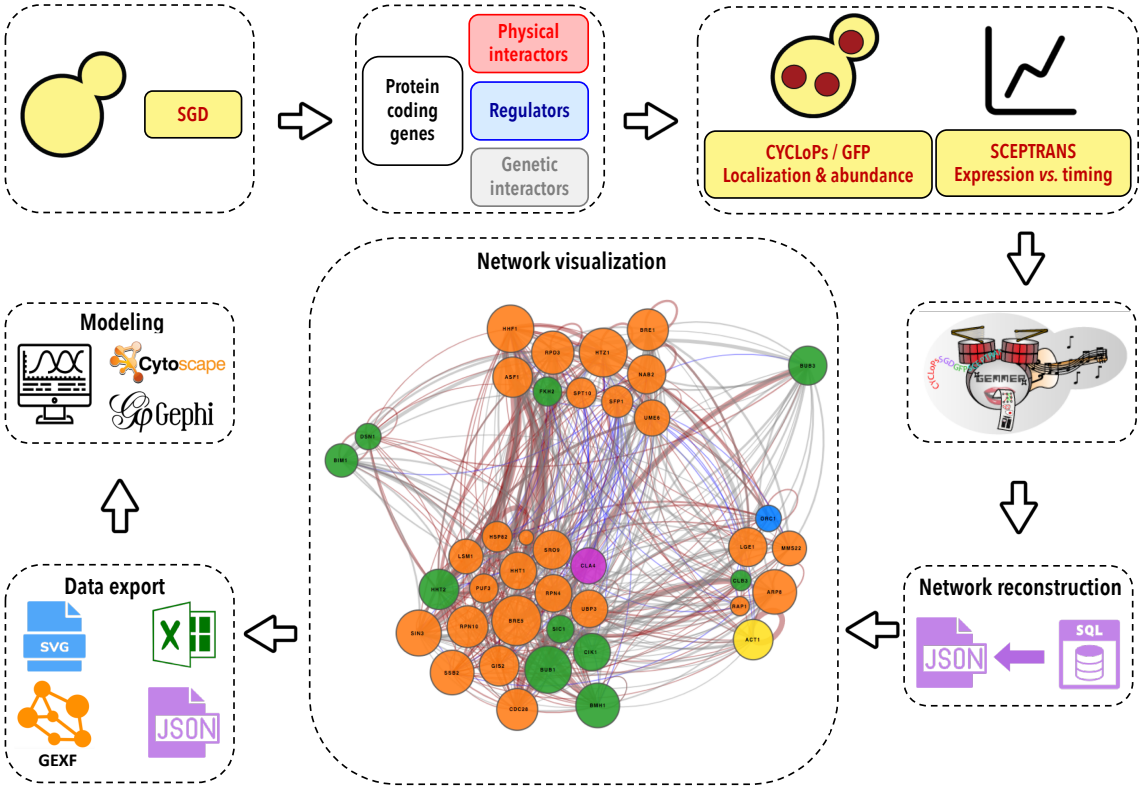


Figure 4.1: GEMMER workflow. GEMMER integrates data from the SGD, CYCLoPs, GFP and SCEPTRANS databases, and provides a user-friendly web interface to interact with this data. Physical, genetic and regulatory interactions for protein-coding genes and their annotations from SGD are integrated with the localization and abundance data from CYCLoPs and YeastGFP and with the timing and assignment of specific cell cycle phases of peak transcription retrieved from SCEPTRANS. The user may query GEMMER for an interaction network fulfilling certain requirements, and GEMMER interfaces with the SQL database to produce a JSON file representing an interaction network corresponding to those requirements. By using D3.js, Cola.js and Cytoscape.js libraries, GEMMER provides interactive, publication-quality visualizations of the interaction networks that may be exported to SVG, Excel, JSON and GEXF formats. GEMMER has been designed to aid in generation of multi-scale visualizations and models for the budding yeast *Saccharomyces cerevisiae*. To this end, the exported data may be imported into desktop applications such as Cytoscape and Gephi.

and interfaces with the SQLite database, ultimately generating a JSON file of the network to be visualized. GEMMER then visualizes the network as a force-directed graph by using the JavaScript library D3js, which reads the JSON file. In addition, alternative visualizations such as hierarchical edge bundling and a circular layout are provided, together with a constraint-based layout that implements compartment separation with coloured boxes. The latter two make use of Cytoscape.js [14] and Cola.js (`http://marvl.infotech.monash.edu.au/webcola/`), respectively. Accompanying the visualization(s), tables are pro-

vided with information about each protein and interaction within the network as well as links via the PubMed search engine to publications with experimental evidence.

Export options provided are: SVG for the network visualization, JSON and GEXF for the interaction network and Excel workbook for the raw data. The Excel workbook and the GEXF network may be imported into Cytoscape [15] and Gephi [16], respectively, for further analysis and model building. The webpage design utilizes the Bootstrap library, which, together with the universality of the D3js JavaScript library, allows the user to run GEMMER on any of the modern browsers such as Firefox, Google Chrome, and Safari.

GEMMER is freely available at `http://gemmer.barberislab.com`. Source code, written in Python, JavaScript library D3js, PHP and JSON, is freely available at `https://github.com/barberislab/gemmer`.

4.4 Conclusions

GEMMER has been developed to integrate existing data on proteins in budding yeast, by providing publication-quality visualizations of their interactions. The tool serves as a data-integration hub, and its visualizations should aid exploration and understanding of complex networks encountered in multi-scale models. The currently available data and the implemented features, expandable in the future, achieve this goal. We aim for GEMMER to become a go-to tool in support of the yeast community.

References

[1] A. Conan Doyle. *The Adventures of Sherlock Holmes*. London, United Kingdom: George Newnes, 1892, p. 307.

[2] H. Kitano. "Computational systems biology". *Nature* 420 (2002), pp. 206–210. 10.1038/nature01254.

[3] F. Castiglione *et al.* "Modeling biology spanning different scales: an open challenge." *BioMed research international* 2014 (2014), p. 902545. 10.1155/2014/902545.

[4] J. M. Cherry *et al.* "Saccharomyces Genome Database: the genomics resource of budding yeast". *Nucleic Acids Research* 40 (2012), pp. D700–D705. 10.1093/nar/gkr1029.

[5] J. L. Y. Koh *et al.* "CYCLoPs: A Comprehensive Database Constructed from Automated Analysis of Protein Abundance and Subcellular Localization Patterns in Saccharomyces cerevisiae." *G3 (Bethesda, Md.)* 5 (2015), pp. 1223–32. 10.1534/g3.115.017830.

[6] W.-k. Huh *et al.* "Global analysis of protein localization in budding yeast". *Nature* 425 (2003), pp. 686–691. 10.1038/nature02026.

[7] S. Ghaemmaghami *et al.* "Global analysis of protein expression in yeast". *Nature* 425 (2003), pp. 737–741. 10.1038/nature02046.

[8] M. Rowicka, A. Kudlicki, B. P. Tu, and Z. Otwinowski. "High-resolution timing of cell cycle-regulated gene expression". *Proceedings of the National Academy of Sciences* 104 (2007), pp. 16892–16897. 10.1073/pnas.0706022104.

[9] A. Kudlicki, M. Rowicka, and Z. Otwinowski. "SCEPTRANS: an online tool for analyzing periodic transcription in yeast". *Bioinformatics* 23 (2007), pp. 1559–1561. 10.1093/bioinformatics/btm126.

[10] D. Szklarczyk *et al.* "The STRING database in 2017: quality-controlled protein–protein association networks, made broadly accessible". *Nucleic Acids Research* 45 (2017), pp. D362–D368. 10.1093/nar/gkw937.

[11] A. Chatr-aryamontri *et al.* "The BioGRID interaction database: 2017 update". *Nucleic Acids Research* 45 (2017), pp. D369–D379. 10.1093/nar/gkw1102.

[12] D. Alonso-López *et al.* "APID interactomes: providing proteome-based interactomes with controlled quality for multiple species and derived networks". *Nucleic Acids Research* 44 (2016), W529–W535. 10.1093/nar/gkw363.

[13] S. Orchard *et al.* "The MIntAct project—IntAct as a common curation platform for 11 molecular interaction databases". *Nucleic Acids Research* 42 (2014), pp. D358–D363. 10.1093/nar/gkt1115.

[14] M. Franz *et al.* "Cytoscape.js: a graph theory library for visualisation and analysis." *Bioinformatics (Oxford, England)* 32 (2016), pp. 309–11. 10.1093/bioinformatics/btv557.

[15] P. Shannon *et al.* "Cytoscape: a software environment for integrated models of biomolecular interaction networks." *Genome research* 13 (2003), pp. 2498–504. 10.1101/gr.1239303.

[16] M. Bastian, S. Heymann, and M. Jacomy. *Gephi: An Open Source Software for Exploring and Manipulating Networks*. 2009. 10.1136/qshc.2004.010033.

CHAPTER 5

Gear-shifting across thermodynamic landscapes

5.1 Introduction 141
The coupling of anabolism and catabolism 142
Stochastic fluctuations, attractors and Onsager reciprocity . . . 143
The phenomenological stoichiometry 152
5.2 Results 156
NET works after all? Stability criteria 156
The variomatic strategy 157
Gear-shifting in acetogenic bacteria? 158
The WL pathway and a hydrogenase are essential 160
ATP coupled to the acetogenesis pathway 161
BHB yield coupled to acetogenesis for plastic production 165
Potential for gear shifting in *C. ljungdahlii* 165
Gear-shifting in *S. sulfataricus* 168
5.3 Discussion 169
5.4 Methods 170
Flux balance analysis 170
Extending the *C. ljungdahlii* GeMM by Nagarajan et al. 171
Model checking and visualization 172
Reproducibility 172

Adapted from:
T.D.G.A. Mondeel, S. Rehman, Y. Zhang, M. Verma, P. Dürre, M. Barberis, H. V. Westerhoff, Maps for when the living gets tough: Maneuvering through a hostile energy landscape, IFAC-PapersOnLine. 49 (2016) 364–370. 10.1016/j.ifacol.2017.03.002

A. Abudukelimu, T.D.G.A. Mondeel, M. Barberis, H. V. Westerhoff, Learning to read and write in evolution: from static pseudoenzymes and pseudosignalers to dynamic gear shifters, Biochem. Soc. Trans. 45 (2017) 635–652. 10.1042/BST20160281

T.D.G.A. Mondeel, S. Astrologo, Y. Zhang, H. V. Westerhoff, NET works after all? Engineering robustness through diversity, IFAC-PapersOnLine. 51 (2018). 10.1016/j.ifacol.2018.09.007

> "Microbes make up 80 percent of all biomass, says Carl Woese. In one fifth of a teaspoon of seawater there's a million bacteria (and 10 million viruses), Craig Venter says, adding, "If you don't like bacteria, you're on the wrong planet. This is the planet of the bacteria." That means most of the planet's living metabolism is microbial. When James Lovelock was trying to figure out where the gases come from that make the Earth's atmosphere such an artifact of life (the Gaia Hypothesis), it was microbiologist Lynn Margulis who had the answer for him. Microbes run our atmosphere. They also run much of our body. [...] This biotech century will be microbe enhanced and maybe microbe inspired. [...] Confronting a difficult problem we might fruitfully ask, 'What would a microbe do?"'
>
> — Stewart Brand[1]

Abstract

With the genome sequencing of thousands of organisms, a scaffold has become available for data integration: molecular information can now be organized by attaching it to the genes and their gene-expression products forming maps that enable functional interpretation of the fitness of the genome. Classical thermodynamics restricts such network-engineering. These restrictions are independent of mechanism and kinetics, and thereby inescapable. Forgetting these restrictions can lead to over-optimistic network designs: is every biochemical network design feasible, provided one puts classical thermodynamics in place? Or, are there other, ill-recognized, generic restrictions to bioengineering? We here discuss how processes away from equilibrium must indeed depend on kinetics and mechanism, but, importantly, not on all kinetic and mechanistic details: There are limitations to what the engineering of mechanisms and kinetics can achieve. Importantly, the Non-Equilibrum Thermodynamics (NET) methodology also shows that system properties that are possible, can be engineered only in certain ways. The NET methodology enables understanding and perhaps engineering a performance that, by adjusting the network, remains optimal when conditions are changing. We introduce 'variomatic' gear shifting as a way that some cells may use to self-engineer their ways to maximal growth rates in environments that lack robust resources, such as in environments with fluctuating oxygen levels. Four billion years ago, bioenergetics may have shuffled 'electron-writers', producing various networks that all served the same function of anaerobic ATP synthesis and carbon assimilation from hydrogen and carbon dioxide, but at different ATP/acetate ratios. This would have enabled organisms to deal with variable challenges of energy need and substrate supply. The same principle might enable 'gear-shifting' in real time, by dynamically generating different pseudo-redox enzymes, reshuffling their coenzymes, and rerouting network fluxes. Using in-silico analyses, we show that gear-shifting may indeed occur in *Clostridium ljungdahlii* and *S. solfataricus*. We shall address how *Clostridium ljungdahlii* may use at

[1] In response to the 2011 question posed by Edge.org: What scientific concept would improve everybody's cognitive toolkit? `https://www.edge.org/response-detail/11863`

least two special features and one special pathway to this end: gear-shifting, electron bifurcation and the Wood-Ljungdahl pathway. Additionally, we find that there should be a definite effect of the choices of redox equivalents in the Wood-Ljungdahl pathway and the hydrogenase on the yield of interesting products like hydroxybutyrate.

5.1 Introduction

Molecules are composed of neutrons, protons and electrons. Due to the electric charges of the latter two elementary particles, different molecules have different energies. Energies also differ between different dynamic conformations of the molecules. Usually there is an equilibration between these conformations such that a molecule can be characterized by an average energy. The important corollary is that the impact of the molecule on the performance of any system of interest can be described in terms of that average energy rather than in terms of the impacts of all the molecules of the same identity but in the different conformations. With the myriads of individual molecules in living organisms with millions each of conformational states, the phenomenon that impact can be described in terms of averages is essential for both bioscience and bioengineering. Not even the fastest computer will ever be able to compute the behaviors of all the individual molecules of a living cell, first because its capacity is too small, and second because information is lacking on the initial state of all the individual molecules: we would not even know where to start computing.

However, we are not interested in the behavior of every individual molecule of a living cell. We are usually interested in the behavior of populations of cells that perform a certain function either in the sense of biotechnology or in the sense of pathophysiology. Accordingly, understanding the behavior of populations of cells as a function of the *average* properties of all the molecules of a given identity within them, is close enough to what we really want. With the added acknowledgement that the molecules of the various identities are engaged in dynamic networking that through nonlinear interactions gives rise to new functional properties, this understanding is the ambition of systems biology. The ambition to predict and engineer towards a useful behavior is then the ambition of systems bioengineering.

Molecular mechanics or molecular dynamics is the discipline that studies the dynamic behavior of individual molecules. Statistical mechanics deals with the statistical properties of ensembles of such molecules. It argues in terms of probabilities and probability distributions. Whenever the average behavior of an ensemble of molecules can be described in terms of averaged properties such as averages, variances and skewness, the discipline becomes statistical thermodynamics, and when averages suffice, generalized thermodynamics is the discipline in charge. What is commonly called 'kinetics' discusses reaction rates in terms of ensemble averaged concentrations or, if more sophisticated, in terms of activities. In this sense it is a branch of generalized thermodynamics. In practice kinetics also has an empirical or a quasi-probabilistic basis and in its extrapolations it is

not necessarily weary of limitations imposed by thermodynamic principles.

Equilibrium thermodynamics champions at least two such principles. One, which it shares with (quantum) mechanics, is the law of conservation of energy (U). Energy U can be brought into a system through heat import, by doing work on the system, or by importing substances with high energy content [1]. It cannot be produced or annihilated (dissipated) however. The other is the law of entropy production, which states that entropy (S) can only be produced and not consumed or destroyed. Entropy is the logarithm of the number of realizations of a system multiplied by the Boltzmann constant. The second law of thermodynamics basically rewords a probabilistic law i.e. *ceteris paribus* (i.e. in splendid isolation) a system will move from a state with lower probability to a state with higher probability, whenever such movement is possible, and not in the opposite direction. Movement in this opposite direction would destroy entropy. Ordered states usually have a smaller multiplicity and hence a lower probability and entropy, than chaotic states of a system at the same energy. Hence this second law maintains that systems in splendid isolation cannot become more ordered and the first law states that they cannot grow from low to high energy content. The paradox that the development of an adult organism from a fertilized egg should then be impossible, is resolved by acknowledging that such developing living systems must be open, in order to import energy and to export more entropy than the order (negative entropy) they create internally. For open 'metabolic' systems the two laws of thermodynamics reduce to the requirement that 'metabolic' (approximately equal to Gibbs') free energy can only be dissipated and must in fact be dissipated to maintain and proliferate the living state [1]. At equilibrium the free energy differences of all reactions that are possible should equal zero. Autonomous reactions, i.e. reactions not coupled to any other processes cannot run uphill in terms of the free energy. This second law of thermodynamics is general, i.e. independent of mechanism. No enzyme or network mechanism can be engineered so that it circumvents this limitation: equilibrium thermodynamics has the strength that it is completely independent of mechanism.

An underlying and often overlooked limitation is however that the validity of this thermodynamics itself and therewith the validity of its second law, depends on the *proviso* mentioned above that dynamic behavior can be described in terms of average concentrations. In this chapter we shall effectively demonstrate that if that proviso is not met, the second law *per se* may not be valid for some important biological systems.

The coupling of anabolism and catabolism

In its first incarnation, non-equilibrium thermodynamics (NET) dealt with the paradox how Gibbs energy could be dissipated yet increase at the same time. We shall envisage two processes in terms of fluxes J_a and J_c, positive when proceeding in the forward direction, i.e. from substrates to products (which for growth is biomass), each associated with a ΔG equal to the Gibbs energy of the products minus the Gibbs energy of the reactants. The two processes represent growth (or anabolism) and catabolism, as indicated by subscripts a and c, respectively. Note

that a negative value of ΔG indicates a thermodynamically favorable reaction through which free energy is dissipated.

Assuming that there is only growth, the rate of Gibbs energy dissipation (Φ) should equal [1]

$$\Phi \stackrel{\text{def}}{=} -\frac{d_i G}{\mathrm{dt}} = J_a \cdot -\Delta G_a > 0. \tag{5.1}$$

$\frac{d_i G}{dt}$ is (an incomplete differential) equal to the Gibbs energy increase due to Gibbs energy production, which has to be negative according to the second law of thermodynamics [2]. Because ΔG_a is positive (usually, though not always [1]), J_a must be negative, implying that growth of a microorganism, or analogously, the production of value added compounds, should be impossible (in fact negative, corresponding to death) according to this equation alone. This leads to the paradox that life cannot exist (or persist) although it does.

The resolution to this paradox that is practiced emphatically by living systems is that the thermodynamically uphill and thereby forbidden anabolic process is coupled to a thermodynamically downhill process, often called catabolism (referred to by subscript c), at positive flux J_c that dissipates more free energy than the anabolic process consumes when J_a is positive. Consequently, in total the Gibbs energy is then dissipated at a positive rate Φ:

$$\Phi = J_a \cdot -\Delta G_a + J_c \cdot -\Delta G_c > 0, \tag{5.2}$$

under the condition that

$$J_c \cdot -\Delta G_c > J_a \cdot \Delta G_a.$$

The coupling does not occur automatically however. It requires some coupling mechanism by which the anabolic flux is pushed towards biosynthesis by the thermodynamic driving force provided by the free energy of catabolism. Thereby the anabolic flux becomes a function of both free energy differences. Accepting that this may be so for both fluxes, expanding both functions as Taylor series around equilibrium, using that at zero free energy differences the fluxes must be zero, and neglecting all higher than first order terms, the coupling can be described by *phenomenological coefficients* L_{ac} and L_{ca} in the phenomenological flow-force relations:

$$\begin{aligned} J_a &= L_{aa} \cdot -\Delta G_a + L_{ac} \cdot -\Delta G_c \\ J_c &= L_{ca} \cdot -\Delta G_a + L_{cc} \cdot -\Delta G_c, \end{aligned} \tag{5.3}$$

where ΔG_a, $-\Delta G_c$, L_{aa}, L_{cc}, L_{ca}, L_{ac} and J_c should be positive, so that also J_a becomes positive, implying the occurrence of growth (see below).

Stochastic fluctuations, attractors and Onsager reciprocity

Eq. 5.3 describe the deterministic behavior of the average system in terms of average Gibbs energy differences. The molecular world is more variable than this deterministic behavior however. It is subject to quasi-random reaction events

that may transiently violate the second law of thermodynamics. Such transitions must be followed by transitions that return the overall behavior of the system to the deterministic behavior. The underlying reason for this is that although systems tend to move from a low to a higher probability state, this probabilistic law is itself subject to stochasticity: systems can transiently move to a less probable state; the law of movement towards higher probability is only true on average. Both the deterministic behavior and the behavior that is in transient violation of deterministic behavior are due to the same processes of rapid energy exchange between molecules and their environments that occurs at temperatures above zero Kelvin. In this sense the beautiful figures in cell-biology textbooks fall short of the reality that is much more chaotic. Only the average could behave in accordance with the diagrams.

At some initial time point, any real system may be in any state of any probability [3]. With time it will then, on average, move to more probable states and as time progresses it will be 'caught' by an environment of states that are highly probable, i.e., by a so-called attractor. These attractors need not be the most probable states but they should be situated on a hill in the probability landscape, surrounded on most sides by states of lower probability. Stable steady states are such attractors. In such steady states all concentrations are often said to be independent of time, but in reality they are not precisely so: They still fluctuate and are thereby varying with time. It is their time average over some limited time span that is independent of time. Both thermodynamics and kinetics deal with such time averaged fluctuating concentrations and they may do this even outside steady states, as many fluctuations are faster than the times characteristic of the evolution of the system.

When systems are not yet close to an attractor, there are often great differences between individual systems, making such conditions unattractive and unuseful for scientific analysis or engineering. Engineering only one out of every one million cells in a population in terms of producing something useful is not usually relevant for bioengineering, because the corresponding productivity will be low. Existing methods of kinetics and non-equilibrium thermodynamics therefore only address systems that are already in densely-populated attractor states or attractor trajectories. Because of the closeness of those systems and because of the property that they tend to remain close to the attractor, the fluctuations around the attractor state are regular [3] (in the sense that they constitute relatively narrow Gaussian distributions).

The fact that most observable and relevant systems are stably in attractor states, leads to an important law: the system under consideration is stable towards all actual fluctuations. Refining the definition of stability in the sense of Lyapunov, this means that after any possible fluctuation, the system will on average ultimately return to a state that is infinitesimally close to the attractor itself [1]. The return to the attractor state after the fluctuation follows deterministic behavior; or perhaps rather vice versa: the deterministic behavior follows the same path as the response to a fluctuation [4].

The equilibrium state is an attractor. In the above described example, equilibrium is where both ΔG are equal to zero and therefore $\Phi_{eq} = 0$. Considering

fluctuations, or corresponding variations in the free energy of catabolism only, to be indicated by $\delta\Delta G_c$, the excess free-energy dissipation *subsequent* to the fluctuation is:

$$\delta\Phi = \delta J_c \cdot \delta(-\Delta G_c) = L_{cc} \cdot (\delta\Delta G_c)^2 . \tag{5.4}$$

δJ_c represents the catabolic flux that arises immediately after the fluctuation (or variation) in the Gibbs energy of catabolism, i.e. as a result thereof. Subsequent to the fluctuation or variation the system must return to the attractor as a deterministic process and must therefore have a positive free energy dissipation, i.e.:

$$\Phi_{eq} + \delta\Phi > 0.$$

Because the fluctuation or variation started at equilibrium where free energy dissipation Φ was zero, the excess free energy dissipation must be positive on average

$$\delta\Phi > 0.$$

Accordingly, given eq. 5.4, the second law of thermodynamics implies that the phenomenological coefficient L_{cc} must be positive.

The same argument requires that L_{aa} be positive and sheds light on a paradox provoked by the phenomenological equations (5.3): In the absence of coupling of anabolism to catabolism, i.e. when the cross coefficient $L_{ac} = 0$, there can be no growth, i.e. J_a must be negative as ΔG_a is (usually [1]) positive. Growth thereby depends on a positive cross coefficient L_{ac}. Assuming that the magnitude of this cross coefficient is subject to catalysis, one might propose to add a high amount of the corresponding catalyst and thereby obtain as much growth as one would want. Could such a miraculous growth machine exist?

The answer is negative because the cross coefficients are limited in magnitude:

$$L_{ac} + L_{ca} < 2 \cdot \sqrt{L_{aa} \cdot L_{cc}}.$$

This is because also after any combined fluctuation in the two free energies the free energy dissipation should be positive. Writing such fluctuations as $\delta\Delta G_a$ and $\delta\Delta G_c$ and using the phenomenological flow-force relations, ones finds for the Gibbs energy dissipation after the fluctuation:

$$\begin{aligned} 0 &< \delta\Phi \\ &= \left(\sqrt{L_{aa}} \cdot \delta\Delta G_a + \sqrt{L_{cc}} \cdot \delta\Delta G_c\right)^2 \\ &+ \left(L_{ac} + L_{ca} - 2 \cdot \sqrt{L_{aa} \cdot L_{cc}}\right) \cdot \delta\Delta G_a \cdot \delta\Delta G_c \end{aligned}$$

We define y and the degree of coupling q by:

$$y \stackrel{\text{def}}{=} \frac{\sqrt{L_{aa}} \cdot \delta\Delta G_a}{\sqrt{L_{cc}} \cdot \delta\Delta G_c}$$

$$q \stackrel{\text{def}}{=} \frac{L_{ac} + L_{ca}}{2 \cdot \sqrt{L_{aa} \cdot L_{cc}}},$$

so that the free energy dissipation becomes a parabolic function of y:

$$0 < \frac{\delta\Phi}{L_{cc} \cdot (\delta\Delta G_c)^2} = (y+1)^2 + 2 \cdot (q-1) \cdot y.$$

Because the free energy fluctuations can be arbitrary, also y can assume any real value. The parabolic function however is always positive for $q^2 < 1$. A necessary condition for this is that:

$$|L_{ac}| \leq 2 \cdot \sqrt{L_{aa} \cdot L_{cc}}$$

This then proves that the cross coefficient L_{ac} is limited in magnitude and that the miraculous growth machine above proposed is in conflict with the second law of thermodynamics.

More generally, the matrix $\mathbf{L}$ of phenomenological coefficients should be positive definite, as for any set of values of the thermodynamic forces:

$$G^T \cdot \mathbf{L} \cdot G = \begin{pmatrix} G_a & G_c \end{pmatrix} \cdot \begin{pmatrix} L_{aa} & L_{ca} \\ L_{ac} & L_{cc} \end{pmatrix} \cdot \begin{pmatrix} G_a \\ G_c \end{pmatrix} > 0$$

Where the middle part of the equation exemplifies for the two-dimensional case.

Fluctuations versus deterministic behavior

It was Onsager [5] who introduced the principle that we employed above when mentioning fluctuations, i.e. the principle that the deterministic behavior of a system initially displaced somewhat from equilibrium due to a perturbation, should be equal to the system's *average* behavior when relaxing back through fluctuations from that same initial state. This should also be so if that initial state had arisen through a fluctuation (see equation 3.1 in [4]). The time average in equation 1.1 of [4] is over an infinitely long period over which the initial displacement may occur spontaneously through a fluctuation, implying that the principle addresses the average behavior of lots of occurrences, which is what we here call the deterministic behavior. This principle may seem obvious because the deterministic autonomous behavior of a system can only be due to fluctuations; because there is no other 'force', it is the fluctuations that are responsible for any dynamics of the system after a transient perturbation or fluctuation, and this includes the average dynamics, which is what we observe as the deterministic behavior [1].

As this principle seems obvious, it is worth our while to examine when it should fail. It should fail if the system were subject to persistent external forces differentiating between the deterministic and the fluctuation case, or if the system has a memory of how it arrived at its non-equilibrium state, i.e. if it is non Markovian. If in the deterministic setting an active transcription state were achieved by a chromatin modification that persists but is not explicit in the description of the system's state, whereas in the fluctuation setting it was achieved by a mere structural fluctuation induced by Brownian motion energy, the principle might fail.

Onsager reciprocity

The catabolic reaction may consist of the breakdown of a substrate S to a product P, so that:

$$\Delta G_c = \mu_P - \mu_S,$$

where μ_S and μ_P refer to the chemical potential of catabolic substrate S and catabolic product P respectively. Eq. 5.3 states that around equilibrium the catabolic flux is proportional to this free energy drop of catabolism. This implies that the dependence of the flux on the chemical potential of the substrate is equal to minus its dependence on the chemical potential of the product. Assuming that this is not the case, would lead to the equation

$$J_c = L_{cc,P} \cdot \mu_P - L_{cc,S} \cdot \mu_S,$$

with $L_{cc,P}$ differing from $L_{cc,S}$. In an open network, S and P can fluctuate independently and also in such a way that their difference is not affected by the fluctuation. For such an equal fluctuation ($\delta\mu_S$) in the chemical potentials of the substrate and the product one finds

$$\delta J_c = (L_{cc,P} - L_{cc,S}) \cdot \delta\mu_S.$$

Because ΔG_c has remained equal to zero, there is no driving force for a catabolic flux in either direction, so that J_c must remain zero. Hence for a non-zero $\delta\mu_S$:

$$L_{cc,P} = L_{cc,S} \stackrel{\text{def}}{=} L_{cc}. \tag{5.5}$$

This confirms that one can write the two fluxes in eq. 5.3 as linear functions of the free energy differences rather than of the individual chemical potentials. Eq. 5.5 thereby serves as an example of Onsager's reciprocity relations [5].

The anabolic reaction may consist of the conversion of substrate for anabolism A into biomass B, so that the fully coupled reaction between anabolism and catabolism is

$$S + A \rightleftharpoons P + B,$$

with fluxes:

$$J_c = J_a = L \cdot -\Delta G_{tot} = L \cdot (-\Delta G_c - \Delta G_a)$$

For the same reasons as above we can write the flux as a linear function of the overall free energy difference rather than of its individual components. Hence:

$$L_{ca} = -\frac{\partial J_c}{\partial \Delta G_a} = -L = -\frac{\partial J_a}{\partial \Delta G_c} = L_{ac},$$

which, more than the eq. 5.5, is known as Onsager reciprocity. If there are additional catabolic and anabolic processes that are not coupled to anabolism and catabolism, respectively, then these should be added to the equations. This will not affect the Onsager reciprocity [1].

The Onsager reciprocity discussed above, is valid close to equilibrium. Further away from equilibrium the proof for it breaks down, as:

$$\begin{aligned} 0 < \Phi + \delta\Phi &= (J_c + \delta J_c) \cdot -(\Delta G_c + \delta\Delta G_c) \\ &= -J_c \cdot \Delta G_c - \delta J_c \cdot \Delta G_c - J_c \cdot \delta\Delta G_c - \delta J_c \cdot \delta\Delta G_c \\ &= -J_c \cdot \Delta G_c + L_{cc} \cdot \Delta G_c \cdot \delta\Delta G_c - J_c \cdot \delta\Delta G_c + L_{\mathrm{cc}} \cdot (\delta\Delta G_c)^2 \end{aligned}$$

In this case there is no reason for L_{cc} to be positive, as the leading term $-J_c \cdot \Delta G_c$ already guarantees a non-infinitesimal positive free energy dissipation whilst the other terms are infinitesimal. Indeed, Onsager reciprocity has been shown to be absent in actual cases of mitochondrial oxidative phosphorylation, where the phosphorylation flux hardly depended on the free energy of respiration whilst the respiratory flux was greatly reduced by increased phosphorylation potential [1]. Yet the coupling, i.e. the positivity of L_{ca} and of, though perhaps less quantitatively so, L_{ac} persists qualitatively.

Returning to near equilibrium conditions, Onsager reciprocity is also an example of non-equilibrium thermodynamics (NET), in the sense that it (i) addresses systems that are not at equilibrium, (ii) describes the system in fewer than the total number of independent variables (i.e. ΔG_c rather than the individual chemical potentials with as corollary Onsager reciprocity) and (iii) omits some mechanistic detail such as the precise way the coupling (positivity of L_{ca}) is achieved. Yet this NET differs from equilibrium thermodynamics in that it admits some mechanistic detail, i.e. the phenomenon and extent of coupling; the statement that there must be a mechanism making $L_{ac} > 0$.

Microscopic reversibility

As shown in [1] and exemplified above Onsager symmetry is maintained if reaction systems consist of a sum of chemical reactions in which a defined set of substrates is converted to a defined set of products, each at a well-defined reaction stoichiometry. Although in the first of his two 1931 papers on the reciprocal relations [5], Onsager did use chemical reactions to illustrate the principle of detailed balance, he did not use this additive property of chemical reaction systems. Instead he generalized the principle of detailed balance (or 'microscopic reversibility') from a similar principle in chemistry. In the proof of Onsager reciprocity that we formulated above, this detailed balance principle was used only implicitly, i.e. when we observed that because ΔJ_c must be zero $L_{cc,P}$ must equal $L_{cc,S}$. We there implicitly assumed that there was no other process converting S to P running at a rate $\Delta J'_c$ such that the $\Delta J_{c,\mathrm{total}} = \Delta J_c + \Delta J'_c$ was zero with $L_{cc,P}$ differing from $L_{cc,S}$, whereas in realistic biochemical networks there could well be such a parallel process. Needed here is indeed *detailed* balance, i.e. the phenomenon that for every individual process the net Gibbs energy dissipation should equal zero.

This has two corollaries. The first is that at equilibrium no futile cycle of the type $A \rightarrow B \rightarrow C \rightarrow A$ can occur, even though this would not decrease the Gibbs energy of the system and would not seem to violate the second law

of thermodynamics. The equilibrium condition does not only require that the system is minimal in terms of Gibbs energy such that the latter cannot decrease and therefore remains constant. It also requires that all net fluxes through all processes equal zero.

Here an equilibrium system differs in an important sense from a system at steady state. In the latter, the Gibbs energy of the system is constant, but at least some net fluxes differ from zero and Gibbs energy is being dissipated. (It may be noted that mathematicians often use the word 'equilibrium' for steady state, as the fluxes balance ('equilibrate') such that the net flux into any node of the system becomes equal to zero. We will discriminate however between non-equilibrium steady states and equilibrium states, as the fluxes in the former is what makes life live.).

The second corollary is that already here a mechanistic aspect enters non equilibrium thermodynamics: it does not only suffice to know that in a system a substance A can be converted to a substance B, but also whether or not the conversion can be executed by one, or by more actual molecular processes. Suppose one observes that in the reference state there is no net flux converting A to B and then infers that the flux from A to B following a minor increase in the concentration of A (and a concomitant decrease in the concentration of B) should depend as much on the chemical potential of A as on minus the chemical potential of B. For this inference to be valid, one needs to know that there is no additional mechanism where the reaction from A to B is coupled to a second process that is not neutral in terms of Gibbs free energy.

But where does the phenomenon of microscopic reversibility come from? Onsager [5] presented it as a principle adhered to anyway by chemists. The principle requires that for the system $A \leftrightarrow B \leftrightarrow C \leftrightarrow A$, at equilibrium (for instance) there should not only be a balance for every metabolite, e.g.:

$$0 = \frac{dA}{dt} = (k_{AB} + k_{AC}) \cdot A - k_{BA} \cdot B - k_{CA} \cdot C$$

but also a balance between the forward and the reverse flux through any process between any two metabolites, e.g.:

$$0 = k_{AB} \cdot A - k_{BA} \cdot B.$$

Let us consider a deterministic situation where there is no such detailed balance, whilst all concentrations are time independent. Then at equilibrium there should be direct fluxes between each combination of the three compounds A, B, and C, and these three fluxes should be equal to one another:

$$\varphi \stackrel{\text{def}}{=} \varphi_{AB} \stackrel{\text{def}}{=} u_{AB} - u_{BA} = \varphi_{BC} \stackrel{\text{def}}{=} u_{BC} - u_{CB} = \varphi_{CA} \stackrel{\text{def}}{=} u_{CA} - u_{AB} > 0.$$

(We assume that the phantom flux φ runs in the direction $A \to B \to C$. If not then the arguments should be redressed accordingly). The u's refer to the direct unidirectional rates of the reactions. In case of first order reactions, the corresponding

kinetic equations are:

$$\begin{aligned}\varphi &\stackrel{\text{def}}{=} \varphi_{AB} \stackrel{\text{def}}{=} k_{AB} \cdot A - k_{BA} \cdot B \\ &= \varphi_{BC} \stackrel{\text{def}}{=} k_{BC} \cdot B - k_{CB} \cdot C \\ &= \varphi_{CA} \stackrel{\text{def}}{=} k_{CA} \cdot C - k_{AB} \cdot A \\ &> 0.\end{aligned}$$

If they are not already present, we now add specific catalysts, such as enzymes, one for each of these reactions. These catalysts will be highly active but only present at concentrations much lower than those of the molecules A, B, and C. Consequently, they cannot alter the chemical potentials (molar Gibbs energies, related to activities and concentrations) of the three compounds. The equilibrium concentrations of the metabolites A, B, nor C should be affected therefore. A phantom flux may or may not increase in magnitude, but in the consequent equilibrium the three phantom fluxes should again equal each other. The three reactions in this network are independent of each other and also their catalysts can be manipulated independently. Accordingly we now inhibit the catalyst of the reaction $B \leftrightarrow C$ so as to reduce the corresponding flux by 50% with the instantaneous effect that:

$$\varphi_{BC} = \varphi_{AB}/2 = \varphi_{CA}/2$$

and consequently:

$$\frac{dB}{dt} = -\frac{dC}{dt} = \frac{\varphi_{BC}}{2} > 0 = \frac{dA}{dt}$$

and

$$\frac{d\Delta\left(-G_{BC}/(R \cdot T)\right)}{dt} = \frac{\varphi_{BC}}{2} \cdot \left(\frac{1}{B} + \frac{1}{C}\right) > 0.$$

Here we have used the usual expression for the chemical potentials for B and C, e.g.:

$$\mu_B \stackrel{\text{def}}{=} \left(\frac{\partial G}{\partial n_B}\right)_{P,T} = \mu_B^0 + R \cdot T \cdot \ln\left(B/V\right)$$

More in general one would use the fact that the dependence of chemical potential on concentration is positive because the dependence of specific molar entropy on concentration is always negative; with increasing concentration of a solute, order increases [1].

The result implies that whilst the Gibbs energy of the $C \rightarrow B$ reaction started out as zero, it now increases with time to positive magnitudes. Even though the increase in the concentration of B and the decrease in the concentration of C may cause a flux from B to C through A, this flux requires the increase in that Gibbs free energy and can thereby not prevent it.

Now one could add a catalyst of a new reaction that converts B to C whilst performing work on some external system, or synthesizing ATP from ADP and phosphate. This reaction should harvest some of the Gibbs energy in B with respect to C and convert some of that to an external Gibbs energy or to Gibbs energy

in the form of ATP. The result would be a *perpetuum mobile*, i.e. a machine performing net mechanical or chemical work, without any Gibbs energy input. *Perpetua mobilia* violate the second law of thermodynamics and thereby the underlying statistical mechanics principle that autonomous systems never move from high to lower probability [1, 3]. The implication is that the phantom flux should be zero and that detailed balance should indeed apply due the second law of thermodynamics, which in turns stems from the definition of probability in statistical thermodynamics.

Gibbs energy dissipation as driver

Indeed, in a usual formulation of the second law of thermodynamics the Gibbs energy is thereby at a minimum, so that there is no reaction left (within the reaction possibilities) that could lead to dissipation of Gibbs energy. In view of the above, a better formulation is that for any deterministic (and isothermal, isobaric) process to proceed, Gibbs energy needs to be dissipated at a positive rate Φ. We here exemplify this for the chemical conversion of substance X to substance Y which has progressed to the extent ξ. The rate of the reaction is:

$$J_{X\to Y} \stackrel{\text{def}}{=} \frac{d\xi}{dt}$$

What is usually called the Gibbs free energy difference of reaction is then:

$$\Delta G_{X\to Y} \stackrel{\text{def}}{=} \frac{dG}{d\xi} = \mu_Y - \mu_X$$

The second law of thermodynamics requires that the dissipation (loss) rate of Gibbs energy

$$\Phi \stackrel{\text{def}}{=} -\frac{dG}{dt} = J_{X\to Y} \cdot -\Delta G_{X\to Y} > 0$$

hence

$$\frac{-\Delta G_{X\to Y}}{J_{X\to Y}} \stackrel{\text{def}}{=} \frac{1}{L_{XY}} \stackrel{\text{def}}{=} R_{XY} > 0.$$

In words, a reaction cannot have a zero resistance or an infinite conductance, i.e. a net reaction cannot proceed if there is no net driving force for it. The Gibbs energy dissipation Φ, which equals the entropy production [1], keeps processes running, but we identify the Gibbs energy dissipation per unit process, usually called the free energy difference across the process, as the driving force. If successful because there is a nonzero conductance L, this force causes change, i.e. a process at flux J. This generalizes the Newtonian force concept that is associated with the cause of the displacement (when there is lots of friction, or acceleration *in vacuo*) of macroscopic objects in time. The above analysis has the corollary that, as in macroscopic mechanics, systems do not change in the absence of any such a force. Since Biology requires change in the sense of growth and development and growth.

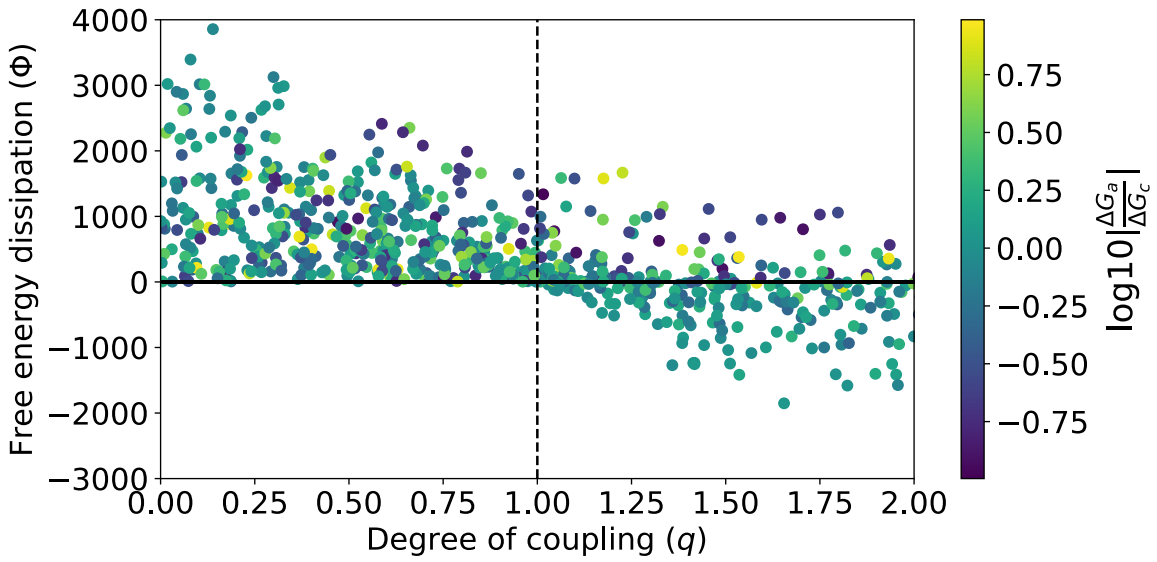


Figure 5.1: Scatterplot to illustrate the (sampled) relationship between Φ and q. 1000 parameter samples for $\Delta G_a \in [0, 40]$, $\Delta G_c \in [-40, 0]$ (with their ratio constrained such that $\log 10 \left| \frac{\Delta G_a}{\Delta G_c} \right| \in [-1, 1]$, i.e. a maximum of ten-fold difference in either direction), and L_{aa}, L_{ac} and L_{cc} constrained to the interval $[0, 2]$ were used to calculate corresponding values for q and Φ. Only for $q < 1$ are Φ values all above 0. This simulation does not constitute a proof but merely serves to illustrate the relationship between Φ and q for the specified ranges of the free energies and the coupling coefficients.

The phenomenological stoichiometry

The degree of coupling [6] is quantified by the ratio of the cross coefficient L_{ac} to the straight coefficients and has been defined, for near equilibrium steady states, as

$$q \stackrel{\text{def}}{=} \frac{L_{ac}}{\sqrt{L_{cc} \cdot L_{aa}}}. \tag{5.6}$$

As illustrated in Fig. 5.1 on the basis of eq. 5.3, this degree of coupling q has to lie between 0 and 1 for it to be guaranteed that the free energy dissipation is positive (a proof is not given here but is available (Westerhoff, unpublished)).

With Z defined as [6]:

$$Z = \sqrt{\frac{L_{aa}}{L_{cc}}}, \tag{5.7}$$

and noting that $q \cdot Z = \frac{L_{ac}}{L_{cc}}$, the so-called flow($J$)-force($-\Delta G$) relations in eq. 5.3 can be rephrased as [1, 6]

$$\begin{aligned} \frac{J_a}{L_{cc} \cdot -\Delta G_c} &= q \cdot Z - Z^2 \cdot \frac{\Delta G_a}{-\Delta G_c} \\ \frac{J_c}{L_{cc} \cdot -\Delta G_c} &= 1 - q \cdot Z \cdot \frac{\Delta G_a}{-\Delta G_c}. \end{aligned} \tag{5.8}$$

These equations show an asymptote when $Z > \left(q \cdot \frac{\Delta G_a}{-\Delta G_c}\right)^{-1}$ (see Fig. 5.2). For q close but not equal to 1, the flow ratio J_a/J_c becomes more substantial with increasing Z until it reaches a maximum well before the asymptote, after which it decreases with Z: then the amount of slippage becomes more and more excessive.

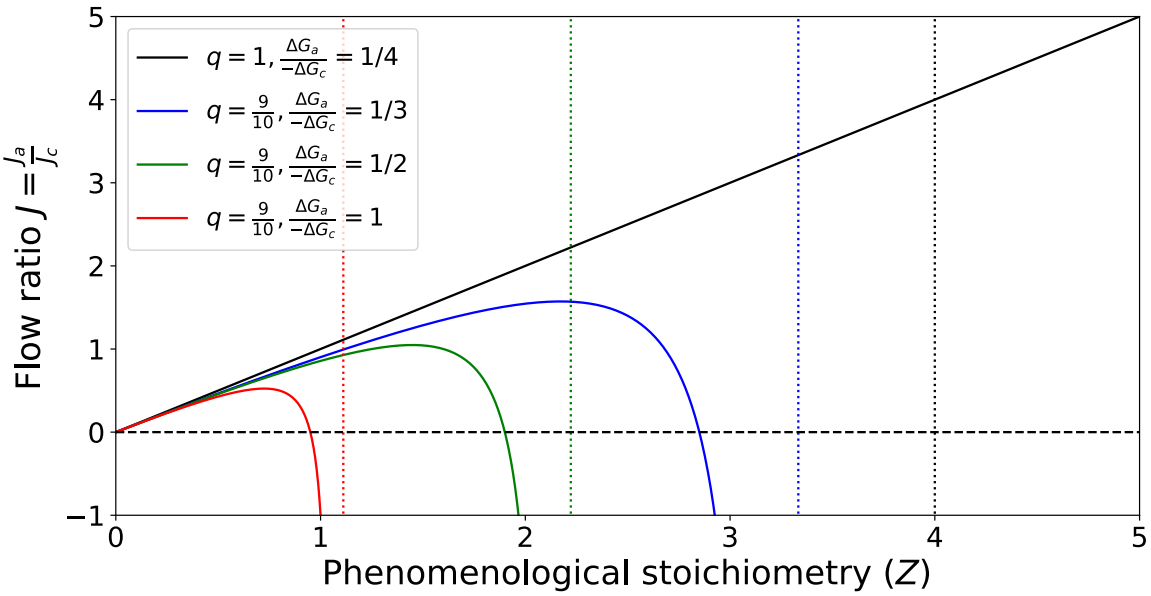


Figure 5.2: Plot of the ratio of anabolic and catabolic fluxes as a function of Z as defined by eq. 5.8. The flow ratio (solid lines) has a vertical asymptote (dotted vertical lines) at $Z = \left(q \cdot \frac{\Delta G_a}{-\Delta G_c}\right)^{-1}$ where $J_c = 0$. In the scenario where $q = 1$ (black line) the flow ratio increases linearly with Z and at the vertical asymptote both fluxes pass through 0. When q is almost but not quite equal to 1 (colored lines) as the phenomenological stoichiometry increases towards the asymptote, the anabolic flux decreases and even becomes negative implying that biomass is being degraded instead of synthesized whilst at the same time catabolism continues though at a slowing rate. At the asymptote the catabolic flux becomes zero and changes sign so that for even higher phenomenological the negative anabolism drives the reversal of catabolism: nutrition is rebuilt driven by the degradation of biomass, i.e. not a highy biological state (not pictured). When the free energy of anabolism is increased the vertical asymptote moves left and the flow ratio decreases (blue vs. green vs. red line).

Anabolism (J_a) decreases linearly with the back pressure exerted by its own free energy (Fig. 5.3). Growth comes to a halt at a 'static head' free energy ratio of:

$$\left(\frac{\Delta G_a}{-\Delta G_c}\right)_{\text{static head}} = \frac{q}{Z} \tag{5.9}$$

Catabolism then still continues (Fig. 5.3) unless there is complete coupling ($q = 1$).

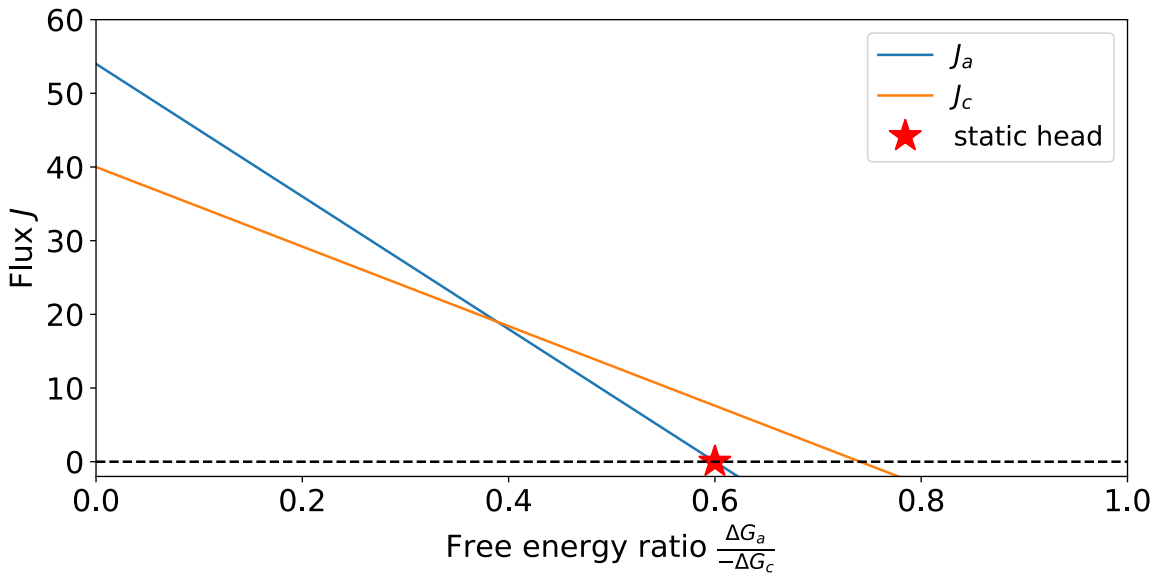


Figure 5.3: Plot of the relationship between the anabolic flux J_a and catabolic flux J_c and the ratio of free energy absorbed and released, respectively, by these processes $\frac{\Delta G_a}{-\Delta G_c}$. For this plot we used $L_{cc} = 1$, $\Delta G_c = -40$, $q = 0.9$ and $Z = 1.5$ which according to eq. 5.9 imply that the static head occurs at $\frac{\Delta G_a}{-\Delta G_c} = 0.6$.

The flow ratio J is given by:

$$J \stackrel{\text{def}}{=} \frac{J_a}{J_c} = Z \cdot \frac{q - Z \cdot \frac{\Delta G_a}{-\Delta G_c}}{1 - q \cdot Z \cdot \frac{\Delta G_a}{-\Delta G_c}} \tag{5.10}$$

This equation, together with Fig. 5.2 (see the black line in Fig. 5.2 or anywhere sufficiently far away from the asymptotes, i.e. in all cases where there is no substantial slippage), explains why Z is called the phenomenological stoichiometry: at full coupling it equals the ratio of growth rate to catabolic flux, i.e. $J = Z$ and at very low free energy of anabolism ($\frac{\Delta G_a}{-\Delta G_c} \approx 0$), it roughly equals the ratio of growth rate to catabolic flux, i.e. $J \approx qZ$, which also equals Z if there is little effective uncoupling. The equation also implies that, unless q is close to 1, at low free energy of anabolism the flow ratio should decrease almost linearly with increasing free energy of anabolism (Fig. 5.4). Because the denominator in this equation goes to zero, at higher such free energies of anabolism, this decrease should become progressively stronger (Fig. 5.4).

The thermodynamic efficiency, equal to the product of the flow ratio with the ratio of free energy differences,

$$\eta = \frac{J_a \cdot \Delta G_a}{J_c \cdot -\Delta G_c} = Z \cdot \frac{q \cdot \frac{\Delta G_a}{-\Delta G_c} - Z \cdot \left(\frac{\Delta G_a}{-\Delta G_c}\right)^2}{1 - q \cdot Z \cdot \frac{\Delta G_a}{-\Delta G_c}} \tag{5.11}$$

thereby exhibits an optimum in its variation with the free energy of anabolism (Fig. 5.4).

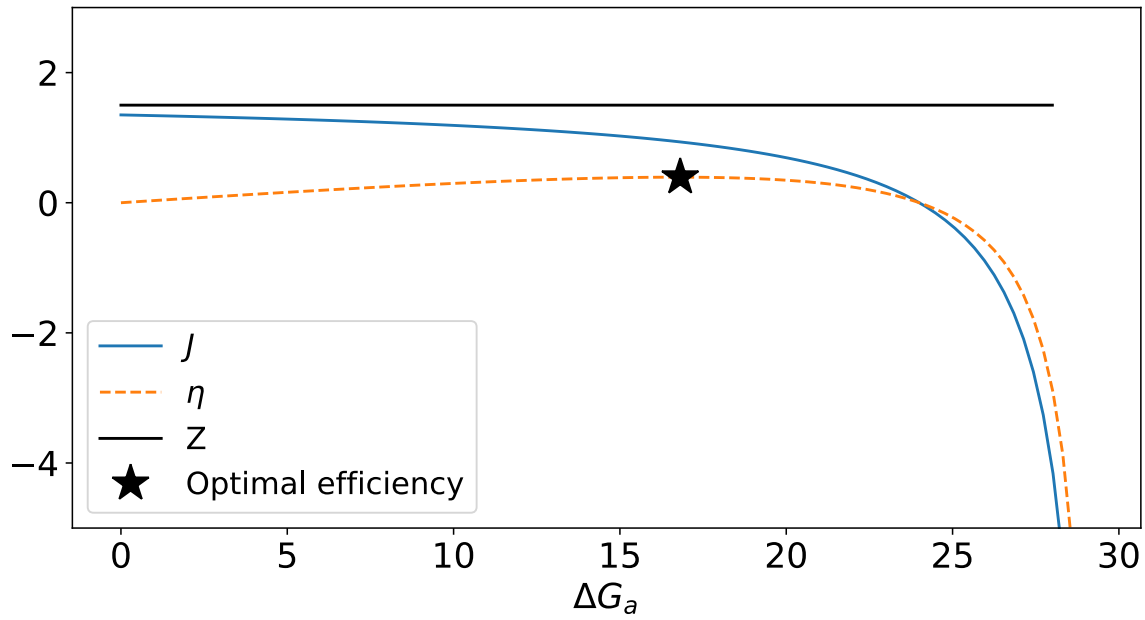


Figure 5.4: Plot of the relationship between the flow ratio and thermodynamic efficiency and the ratio of free energy release. For this plot we used $L_{cc} = 1$, $\Delta G_c = -40$, $q = 0.9$ and $Z = 1.5$.

Above we saw that anabolic flux, flow ratio and efficiency all increase with tighter coupling. After billions of years of evolution one might therefore expect the degree of coupling to equal 1 meaning that coupling would be complete. In reality coupling is less than complete [1]. The reason is of interest to bio-engineering, as q might be a parameter to use in engineering towards better productivity. There are at least three feasible explanations for this lack of complete coupling. The one referring to physical-chemical limitations to stability, is perhaps most pertinent for the many cases where ion-gradient dependent free energy transduction is involved: it may be impossible to make membranes fully tight with respect to ion leakage. A second explanation referring to other free energy dissipating processes that are essential to maintain the living state, so-called maintenance processes, is also feasible. Perhaps a more intriguing explanation was developed by [7]: incomplete coupling might itself be optimal. Determining for each degree of coupling the anabolic free energy optimal for achieving maximum thermodynamic efficiency, and then plotting anabolic flux or flow ratio (all normalized in some way by Z) for varying degrees of coupling as a function of the optimal free energy of anabolism, optima were found at incomplete coupling ([1] page 374). The values found for free energy of anabolism, degree of coupling and efficiency did make sense for mitochondrial oxidative phosphorylation [7] and microbial growth [1]. The success of this theory was surprising because its computations used the above proportional relationships between fluxes and free energy differences, as well as Onsager symmetry, whilst the systems addressed were too far from equilibrium for the proofs of these properties to persist. Anyway the concept that living systems may adjust the degree of coupling in order to

attain optimality, and the idea that there could be more objective functions than growth rate or growth yield, will inspire us below, when we consider adjustment of the phenomenological stoichiometry Z.

In the rest of this chapter, we shall develop a NET approach that should enable us to deal with biological systems which adjust stoichiometries of reactions (or the more closely related phenomenological stoichiometry Z) rather than the degree of coupling. We shall identify a 'variomatic' strategy, i.e. one in which an organism optimizes the stoichiometries. With this we shall show how, useful NET approaches can be developed by allowing a limited amount of mechanistic detail to enter the considerations. We will then illustrate that this variomatic gear-shifting principle may have been adhered to by early life forms on earth: e.g. the acetogenic bacterium *Clostridium ljungdahlii* and the Archaea *S. solfataricus* through in-silico analyses.

5.2 Results

NET works after all? Stability criteria

Metabolic networks exhibit relationships between so-called elasticity coefficients and control coefficients [8, 9]. The most relevant example here is the concentration control connectivity law for systems at steady state [8]

$$\sum_{i=1}^{m} C_i^{X_j} \cdot \varepsilon_{X_k}^i = -\delta_k^j. \tag{5.12}$$

The left hand-side describes the summation over all m reaction processes, of the multiplication of the control exercised by process i over the concentration (chemical potential) of metabolite X_j with the elasticity of process i with respect to the concentration of metabolite X_k. The right–hand side of the equation is minus the Kronecker delta, which equals 0 for $j \neq k$ and 1 for $j = k$. Elasticity coefficients are local derivatives of the logarithm of any process rate with respect to the logarithm of the chemical potential of any freely fluctuating metabolite. They contain the summaries of the kinetic details of the processes that suffice to determine the control of the network. They harbor some kinetic detail but not all. The control coefficients are the dependencies of the logarithm of the chemical potentials at steady state on the logarithm of any of the process activities. Logarithms are here natural ($\ln$), and chemical potentials are normalized by RT. When emphasizing the thermodynamic nature of this law even further, the term 'concentration of' is replaced by 'chemical potential of'. The law has been proven by using the requirement that the deterministic response of the system to a fluctuation in the chemical potential of X_k alone, must be such that that chemical potential returns on average to its initial value, whilst all other chemical potentials remain unchanged [8]. Inverting the argument, we here propose that this connectivity property is the stability criterion for non-equilibrium steady states, also for those beyond the Onsager domain.

The variomatic strategy

When plotting the anabolic flux as a function of the free energy of anabolism (taken relative to the free energy of catabolism, i.e. $\frac{\Delta G_a}{-\Delta G_c}$) for various values of the phenomenological stoichiometry Z, one obtains (Fig. 5.5) a family of downward straight lines, running from $(0, q \cdot Z)$ to $(\frac{q}{Z}, 0)$. At lower values of the free energy of anabolism, higher phenomenological stoichiometries lead to higher anabolic fluxes (green line vs. blue and black lines), but at more challenging free energies of anabolism, the systems with lower values of Z lead to faster anabolism (black line vs. green and blue line). This is akin the effect of shifting to lower gear (decreasing Z) when driving a car (flux) up a steeper and steeper mountain road (higher free energy ratio).

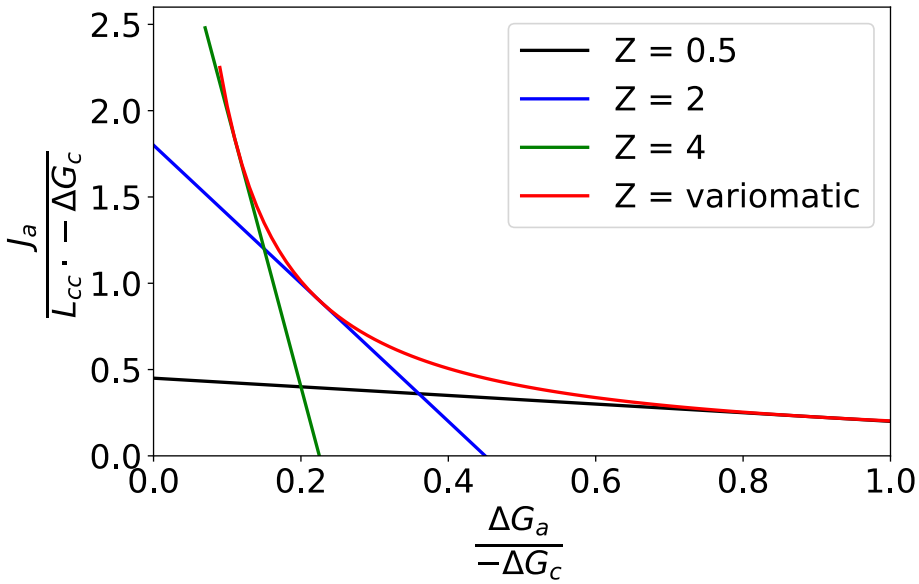


Figure 5.5: Normalized anabolic flux versus ratio of free energy differences. Straight lines: Anabolic flux J_a (normalized by $L_{cc} \cdot -\Delta G_c$) as a function of the free energy of anabolism (normalized by the free energy of catabolism, i.e. $\frac{\Delta G_a}{-\Delta G_c}$), for a degree of coupling $q = 0.9$ at various values (i.e. 0.5, 2 and 4) of the phenomenological stoichiometry Z. The red curve connects the states produced by the so-called variomatic gear shifting defined in the text.

What we here call the variomatic strategy would optimize the gear shifting so that always the highest anabolic flux is attained at every free energy of anabolism. Equating the derivative of the anabolic flux with respect to the stoichiometry, to zero, one finds the optimal phenomenological stoichiometry for every free energy of anabolism:

$$Z_{\text{optimal}} = \frac{1}{2} \cdot q \cdot \frac{-\Delta G_c}{\Delta G_a} \tag{5.13}$$

This confirms that with increasing slope (ΔG_a) it is better to shift to lower gear, i.e. to pathways in the metabolic network that have a reduced phenomenological

stoichiometry. It also shows that if more input free energy is applied, it is better to operate at higher gear. Inserting this expression into the anabolic flow-force relationship for the optimal phenomenological stoichiometry:

$$\frac{J_a}{L_{cc} \cdot -\Delta G_c} = \frac{q^2}{4} \cdot \frac{-\Delta G_c}{\Delta G_a} \tag{5.14}$$

This variomatic curve is shown as the hyperbolic decrease with anabolic free energy in Fig. 5.5. The ratio of the anabolic flux of the fixed stoichiometry network to that of the variomatic network is:

$$\frac{J_{a,\text{ fixed } Z}}{J_{a,\text{ variomatic}}} = 4 \cdot \frac{\Delta G_a}{-\Delta G_c} \cdot \frac{q \cdot Z - Z^2 \cdot \frac{\Delta G_a}{-\Delta G_c}}{q^2} \leq 1 \tag{5.15}$$

which is always smaller than 1 except for when Z precisely corresponds to the optimal gear setting at that anabolic free energy (see above). Variomatic gear shifting in anabolism should be beneficial for anabolic flux, especially under conditions of variable free energies.

When discussing 'gears' we may seem to refer to mechanism, whereas Z is a phenomenological stoichiometry that is only indirectly related to the mechanistic stoichiometries. Moreover, this relationship depends on the mechanism of uncoupling (see page 385 in [1]). Yet, in all mechanisms of uncoupling examined [1], the phenomenological stoichiometry increases monotonically with the mechanistic stoichiometry, if the uncoupling mechanism is constant in activity (see page 385 in [1].

Gear-shifting in acetogenic bacteria?

Our understanding of the origin of life during the early periods of our planet is still lacking [10]. In the early soup of chemicals, some billion years ago, no organic carbon compounds may have been abundant, there was no oxygen [11] and the question of how life got going is therefore intriguing [10]. A critical question in this regard is how the early organisms could produce their ATP or at least extract free energy from their environment and convert this into some utilizable form, and grow autotrophically on the available mixtures of CO, CO_2 and H_2.

C. ljungdahlii is one of the few organisms that are able to grow under the highly challenging conditions that may characterize the early *billenia* of planet earth, in which there was neither organic carbon abundant, nor any other complex source of Gibbs free energy, nor molecular oxygen, or other suitable electron acceptors. Microorganisms called acetogens, like *C. ljungdahlii*, may have been important for the origin of life on this planet [10, 12]. They can fix carbon dioxide anaerobically using hydrogen as the electron donor in processes coupled to the synthesis of ATP [13, 14]. Specifically, *C. ljungdahlii* and other acetogens are able to produce acetate from CO_2 and molecular hydrogen:

$$2\,\mathrm{CO_2} + \ 4\,H_2 \rightarrow \mathrm{CH_3COO^-} + \ \ H^+ + \ 2\,H_2O. \tag{5.16}$$

It has been discussed that acetogens have this fairly unique capability not by possessing a unique protein, but rather by having a unique pathway, i.e. the

Wood-Ljungdahl pathway [15]. We examined whether the information comprised in the genome sequence of *C. ljungdahlii* and the annotation thereof, confirms this capability. To do this, we used the genome-wide metabolic map (GeMM) already published in [16] and applied flux balance analysis (FBA) [17]. We changed medium conditions and the objective function to reflect the acetogenesis problem. We also added any missing reactions that were considered in [15] but not in the published GeMM, see the Methods section 5.4. This involved 5 redox-coenzyme variants of existing reactions in the reconstruction (Table 5.3).

Schuchmann and Müller [15] highlighted the Wood-Ljungdahl pathway (WLP) as a network feature enabling to overcome the difficulties faced by early colonizers of Earth. Since the WLP is ATP-neutral, an additional 'trick' is also required however so as to be enable the organism to harvest the limited amount of Gibbs free energy made available by the WLP. This trick has been identified as electron bifurcation.

Acetogens are home to redox enzymes that interconvert coenzymes such as NAD(H), ferredoxin, a quinone, or NADP(H). These include so-called electron bifurcating enzymes [18], which have three 'writer' domains, enabling them to carry out two different redox reactions starting from the same electron donor but running to two different electron acceptors in the two remaining writer domains, one thermodynamically uphill and the other thermodynamically downhill; the third theoretical reaction should be forbidden by some mechanism: they do this in a coupled way, enabling the first of the three reactions to proceed using the driving force of the second. An example is the non-membrane bound hydrogenase HydABCD of *A. woodii*, which has electron centres (cofactors or prosthetic groups that can store electrons) on board, i.e. iron-sulphur centres and flavins, and three writing domains [19], i.e. for the oxidation of hydrogen, ferredoxin and NADH. By proper coupling of these writings the enzyme can reduce ferredoxin (uphill) at the same time as NAD (downhill) whilst oxidizing hydrogen. Two thirds of the reduced ferredoxin and NAD are next used to reduce carbon dioxide to acetate therewith providing the carbon building blocks necessary for growth biochemistry. One third of the reduced ferredoxin is oxidized by RnfB, a writer in the membrane bound Rnf complex. A second writer in this complex (RnfC) uses the electrons to reduce another molecule of NAD. Possibly also for the earliest bioenergetics of this planet, these writings may have been coupled to the action of a third writer (RnfD), i.e. one that enables the outward movement of sodium ions across the membrane. This generates an electrochemical potential difference for Na^+ across the bacterial membrane, which can then be used by the sodium motive ATPase of the organism to drive the synthesis of ATP [20]. The Rnf complex lacks the cytochromes of the better known, more 'modern', electron transfer chains, and may thereby constitute one of the earliest mechanisms for the generation of an electrochemical potential difference able to drive the synthesis of ATP.

Schuchmann and Müller [15] described how the WLP and electron bifurcation could lead to the production of ATP from ADP and phosphate provided use of two more transmembrane enzyme complexes is made, i.e. the H^+-ATPase and the Rnf complex. They did not show whether other solutions to the ATP synthe-

sis problem may be possible, and whether what they proposed was in immediate concordance with the knowledge integrated through the genome-wide metabolic maps of the organisms [16].

Here, we wish to compare the analysis by Schuchmann and Müller with the predictions of maximal ATP synthesis emanating from the genome-wide metabolic reconstruction of the model acetogen *Clostridium ljungdahlii* through flux-balance analysis (FBA) [17]. Some analysis on the effect of redox equivalents on growth and product synthesis was already present in [16]. Specifically, it was shown that the genome-wide map predicts the possibility of growth on CO_2/H_2 and CO and the effect of various options in redox equivalents were analyzed under the knockout of acetate kinase. We extend on that analysis by including various reactions that were considered in the treatment by Schuchmann and Müller but were not in the GeMM constructed by Nagarajan et al [16]. Specifically, we will investigate the various alternatives in the electron donor-acceptor combinations for various enzymes and their effect on ATP yield coupled to acetogenesis. Additionally, we will focus on the importance of the Wood-Ljungdahl *pathway* as opposed to single enzymes, the need for electron bifurcation and the Nfn complex, the concept of gear-shifting, the requirement of low gear and advantages of high gear operation, and how much product yield might be attained when engineering *C. ljungdahlii* with two additional genes for producing polyhydroxybutyrate (PHB) under various redox alterations.

The WL pathway and a hydrogenase are essential

In line with the analysis in [15] we asked the genome-wide metabolic map to produce acetate rather than biomass at zero maintenance choosing the variant of the formate dehydrogenase that oxidizes ferredoxin. This indeed yields a flux distribution with positive flux through the Wood-Ljungdahl pathway plus the hydrogenase, Nfn and Rnf complexes and the H^+-ATPase. We next deleted every enzyme of this pathway one by one and confirmed that in all cases except the formate dehydrogenase this eliminated the production of acetate. The formate dehydrogenase is not essential because the formate may be alternatively synthesized through the pyruvate formate lyase. This shows that the total genomic information on *C. ljungdahlii* contained in the annotated map, confirms the notion that almost the entire Wood-Ljungdahl *pathway* is essential for autotrophic growth of *C. ljungdahlii* in the presence of hydrogen gas. It is the *pathway* that matters, not *just* a single enzymatic step.

Furthermore, we investigated the essentiality of the hydrogenase. When deleting both variants of the hydrogenase (i.e. the one with Fd + NAD acceptors and the one with Fd + NADP as electron acceptors) the organism was predicted to be unable to make acetate. With either of these hydrogenases the organism was predicted to be able to produce the acid. This shows that the network property allowing for acetate production autotrophically is the presence of the full Wood-Ljungdahl pathway and a hydrogenase.

In *these* computations, we configured the Wood-Ljungdahl pathway so as to use reduced ferredoxin both to reduce carbon dioxide to formate and to reduce

a second such molecule to carbon monoxide. In the former, 'methyl' branch of the pathway, two molecules of NADH were used for the further reduction of the formate to the methyl group. The hydrogenase reduced two oxidized ferredoxin molecules and two NAD molecules by using molecular hydrogen as the electron donor. In these processes one ATP is used in the formyl-THF synthetase reaction, whereas a second one is produced by the reverse operation of acetate kinase. Consequently, the process merely dissipates the 40 kJ/mol of Gibbs energy that the 4 hydrogen molecules plus 2 CO_2 molecules have in excess compared to the acetate molecule [15]; it does not capture part of this Gibbs free energy in the form of phosphorylated ADP. Since some ATP is necessary for the conversion of acetate to biomass according to the biomass synthesis equation, the pathway shown in 5.6 can produce acetate but cannot produce growth. It cannot even take place in realistic organisms that are not growing, since these require Gibbs free energy for maintenance.

ATP coupled to the acetogenesis pathway

In figure 3 of [15] the ATP yield per acetate is considered for 6 different combinations of 2 hydrogenases and 3 forms of the formate dehydrogenase. The chemical reaction 5.16 may be coupled to the generation of up to roughly 1 ATP. The precise yield of ATP per acetate depends on the electron donors and acceptors used in various steps in the WPL and the hydrogenase and on the presence or absence of electron bifurcating steps. Figure 3 in [15] therefore considered the question of finding the unknown coefficient X in

$$2\,CO_2 + 4\,H_2 \rightarrow CH_3COO^- + H^+ + 2\,H_2O + X\,\mathrm{ATP} \qquad (5.17)$$

or at least find a pathway for which X is positive.

It is not trivial that asking the extended genome-wide metabolic map for maximum ATP yield starting from 2 CO_2 + 4 H_2 would yield the same results as the theoretical analysis in [15]. First of all, reactions may be encoded with different stoichiometries and co-factors. Secondly, the complete genome might contain reactions not considered by Schuchmann and Müller. And thirdly, the GeMM constructed by Nagarajan et al. lacked reactions that were considered by these two authors.

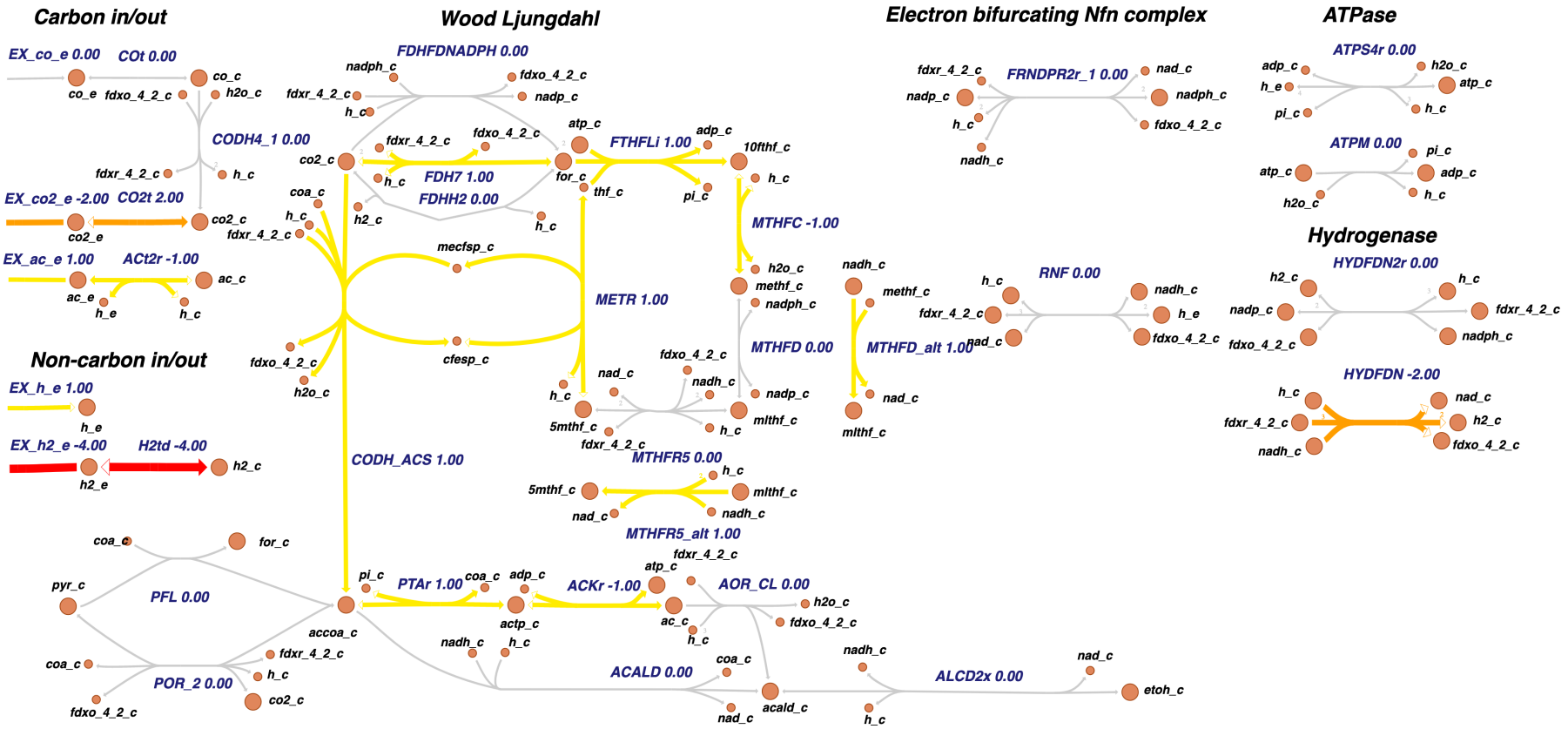


Figure 5.6: Visualization of the flux distribution when asking for ATP and acetate on a CO_2 and H_2 mixture. CO_2 and H_2 are taken up in a 4:2 ratio and together reduce to 1 mole of acetate. Reactions carrying no flux are colored in grey, reactions carrying flux are colored from low (yellow) to high flux (red). The molecule of acetate is exported in symport with a proton, but is not consumed on the inside and is not electrogenic, hence cannot drive ATP synthesis. To reproduce one of the scenarios in the analysis of Schuchmann and Müller, we allowed only the formate dehydrogenase alternative oxidizing ferredoxin, the hydrogenase reducing NAD^+ and the methylene-THF dehydrogenase and reductase as described in [15] to carry flux. The combination of enzyme alternatives shown here does not have any further ability to couple this process to ATP synthesis, as also shown in the 0 (0) containing box in Table 5.1.

As a preliminary check we made sure the model was not capable of generating ATP from nothing in any of the scenarios, see methods section 5.4. We then performed FBA with the maintenance reaction (ATP $\rightarrow$ ADP + phosphate) as the objective function, i.e. requiring the network to phosphorylate ATP. We presented the GeMM with the CO_2/H_2 medium described before and forced the organism to complete the reaction $2\,CO_2 + \,4\,H_2 \rightarrow \,1\,CH_3COO^- + H^+ + 2H_2O$ for the redox reactions that were considered by Schuchmann and Müller but were not in GeMM and compared the results. The corresponding maximal ATP yields are listed in the first two columns of Table 5.1 and the predictions agree perfectly with the ones by Schuchmann and Müller [15].

We extended the scenarios considered in [15] by adding variations of the methylene-THF reductase and the methylene-THF dehydrogenase reactions that were included by Nagarajan et al. [16] (see methods section 5.4). This yields a total number of $3 \times 2 \times 2 \times 2 = 24$ scenarios (Table 5.1). We included the flux through the Nfn complex in parentheses in Table 5.1. With respect to the suggestion by Schumann and Müller that electron bifurcation is essential for ATP production, our results show that if so, this need not be electron bifurcation at the Nfn complex. There were multiple scenarios in which positive ATP yield was obtained whilst the Nfn was inactive (Table 5.1). Also note that the direction of the flux through the Nfn complex flux differs between scenarios.

Judging from the simulation results and from careful manual bookkeeping, the electron bifurcating MTHFR reaction structurally yields an additional 0.5 ATP per acetate. Rather than using a single molecule of NADH to reduce methylene-THF to methyl-THF, it couples this exergonic reaction to the endergonic reduction of ferredoxin by NADH. This electron-bifurcating reaction yields a mole of extra NAD^+ and a mole of reduced ferredoxin per mole of flux through the WLP. Through the Rnf these can together pump 2 additional protons across the membrane which yields 0.5 ATP through the H^+-ATPase. Similarly, the methylene-THF dehydrogenase oxidizing NADH that was envisioned in the analysis by Schuchmann and Müller yields 0.25 ATP more than the NADPH variant which was encoded in the GeMM. In the latter the NADP needs to be re-reduced at the cost of oxidation of NADH costing one half mole of reduced ferredoxin and NADH, which are made available by half a turnover less of the Rnf complex. This results in one proton less being pumped over the membrane by the latter and therefore a 0.25 loss in ATP.

As highlighted by the orange boxes, the FBA computation can violate the second law of thermodynamics, by predicting and ATP/acetate yields > 0.8, the ratio allowed by the energetics of $40kJ$ in the overall process reduction of CO_2 by hydrogen to acetate relative to the Gibbs energy of ATP synthesis from ADP and phosphate which may be close to $48kJ$ per mole [1]. Accordingly, the results in these boxes are unrealistic. The results in the yellow boxes are also unrealistic as they do not yield any ATP for cell maintenance metabolism. The results in the red boxes are unrealistic because they only cost ATP.

	MTHFR	NADH				2 NADH - FD^0			
	MTHFD	NADH		NADPH		NADH		NADPH	
	HYD	FD + NAD	FD + NADP	FD + NAD	FD + NADP	FD + NAD	FD + NADP	FD + NAD	FD + NADP
FDH	FD^{2-}	0 (0)	0.5 (-1)	< 0	0.25 (-0.5)	0.5 (0)	1 (-1)	0.25 (0.5)	0.75 (-0.5)
	FD^{2-} + NADPH	0.125 (0.25)	0.625 (-0.75)	< 0	0.375 (-0.25)	0.625 (0.25)	1.125 (-0.75)	0.375 (0.75)	0.875 (-0.25)
	H_2	0.25 (0)	0.625 (-0.75)	0 (0.5)	0.375 (-0.25)	0.75 (0)	1.125 (-0.75)	0.5 (0.5)	0.875 (-0.25)

Table 5.1: Predicted moles of ATP produced per acetate when performing FBA on the extended genome-wide reconstruction of *C. ljungdahlii* for the Wood-Ljungdahl pathway. We performed the analysis for various electron donors and acceptors, i.e. the donor for the formate dehydrogenase (FDH; for which the electron acceptor is always CO_2, which is reduced to formate), the electron accceptor from the hydrogenase (HYD; its electron donor is always molecular hydrogen), and the electron donor to the methylene-THF reductase (MTHFR; the electron acceptor is always methenyl-THF which is reduced to methylene-THF) and the electron donor to the methyl-THF dehydrogenase (MTHFD; the electron acceptor is always methylene-THF which is reduced to methyl-THF), as indicated by the row and column headers. The electron donor to MTHFR indicated as 2 NADH - Fd^0 refers to an enzyme complex such as the one in *Moorella thermoacetica* [15] where 2 molecules of NADH serve as electron donor and oxidized ferredoxin and CH_2-THF as electron acceptors, which is then another electron bifurcating reaction. The - in front of Fd^0 refers to its use as electron acceptor, not donor. The superscript refers to its formal electric charge. The results in the two left-most columns are in agreement with the analysis by Schuchmann and Müller (their figure 3) and here extended to include the methylene-THF dehydrogenase and reductase alternatives. In parentheses we show the flux through the Nfn complex reaction (relative to the acetate production flux of 1), counted positive when in the direction of oxidizing Fd^{2-} and NADH. Cells colored red have a negative ATP yield and signify it was not feasible to produce acetate. Cells colored orange are produced by the FBA simulations but thermodynamically not feasible. Cells colored yellow correspond to situations in which acetate may be produced but not coupled to ATP production. This similarly cannot occur alone in living organisms due to their maintenance ATP requirement.

	MTHFR	NADH				2 NADH - FD^0			
	MTHFD	NADH		NADPH		NADH		NADPH	
	HYD	FD + NAD	FD + NADP	FD + NAD	FD + NADP	FD + NAD	FD + NADP	FD + NAD	FD + NADP
FDH	FD^{2-}	0.29 (0)	0.47 (-0.65)	0.22 (0.22)	0.35 (-0.24)	0.5 (0.25)	0.5 (-0.13)	0.43 (0.5)	0.5 (-0.13)
	FD^{2-}+NADPH	0.36 (0.25)	0.5 (-0.25)	0.25 (0.38)	0.41 (0.06)	0.5 (0.75)	0.5 (0.12)	0.5 (0.75)	0.5 (0.12)
	H_2	0.43 (0)	0.5 (-0.25)	0.29 (0.29)	0.41 (0.06)	0.5 (0.5)	0.5 (0.13)	0.5 (0.5)	0.5 (0.13)

Table 5.2: Hydroxybutanoyl-CoA yields per acetate under the same scenarios as in Table 5.3 when *C. ljungdahlii* is grown *in silico* on a CO/H_2 mixture in the ratio 2:4. Again, in brackets we highlight the flux through the Nfn reaction.

BHB yield coupled to acetogenesis for plastic production

Betahydroxybutyrate yield coupled to acetogenesis for plastic production *C. ljungdahlii* is currently also investigated with the perspective of producing precursors for biodegradable plastics from waste, for example in the context of the SYNPOL project (`http://www.synpol.org`), by growing the organism on syngas ($CO/CO_2/H_2$) and having it produce beta-hydroxybutyrate, possibly after inserting one or two reaction capabilities into its genome. Genome-wide metabolic modeling, like shown here, may help strain-design for improved yield and growth rates.

Table 5.2 shows a similar scheme to Table 5.1 for the objective of hydroxybutanoyl-CoA synthesis from a CO/H_2 mixture. We conclude that the variants among the WPL enzymes and the hydrogenases and formate dehydrogenases could impact the yield of the product. This suggests that model-assisted genome-editing or over/under-expressing the optimal variants with high yield in Table 5.2, or in vitro selection amongst the variants using an appropriate selective condition could help improve the real-world yield of desired products like hydroxybutyrate and its product poly-hydroxybutyrate (PHB), a biodegradable plastic. For *C. ljungdahlii* this would require the engineering in of an epimerase and a polymerase for the organism to be able to proceed from the beta-butanoyl to the PHB. There may be similar implications for the possible production of other products of possible interest such as butane-2,3 diol and for ethanol as shown in [16].

Potential for gear shifting in *C. ljungdahlii*

The genera Acetobacterium and Clostridium, to which most acetogens belong, are highly versatile in how they bring about the synthesis of acetate coupled to the phosphorylation of ADP. They appear to avail of a redox protein construction kit [21] of enzymes that are similar, except that writer domains [19] of one specificity (e.g. for NAD) have been exchanged with ones of a different specificity (e.g. for Fd). This may enable *Clostridium ljungdahlii* to oxidize hydrogen with either ferredoxin plus NAD or ferredoxin plus NADP as electron acceptor, and to reduce carbon dioxide to formate with ferredoxin, hydrogen, or ferredoxin plus NADPH (1/1) as electron donor. The consequence is that acetogens may function at a variety of yields of ATP per acetate produced (see [13] and Table 5.1).

A formate dehydrogenase that uses ferredoxin rather than hydrogen as electron donor would lack a functional writer that can accept electrons from hydrogen. In terms of its ability to oxidize hydrogen it would appear to be a dead enzyme, a pseudoenzyme indeed [19, 22]. We here propose that many pseudoenzymes are cases of the enzyme construction kit [21] of evolution, where we simply have not yet recognized the writer or reader domain an original reader or writer domain has been replaced by.

The complex formate dehydrogenase of *C. ljungdahlii* that was mentioned above exists in a complex gene cluster (Clju_c06990-07080), but the same organism contains another gene cluster (Clju_c20030–20040) that appears specific for ferredoxin as electron donor. This suggests that the diversity of redox enzymes

with the corresponding diversity in ATP/acetate ratios is not just a diversity of species within genera but may constitute a more dynamic diversity within a single species, such as *C. ljungdahlii*. Such diversity would enable the organism to function at a variety of ATP/acetate production ratios. This would correspond to dynamic gear shifting, which should be useful when the going gets tough for the organism in the sense of an increased ATP/ADP ratio to work against. The organism would then continue to be able to synthesize ATP but at a lower rate by shifting to lower gear.

Given the results in figure 3 of [15] and Table 5.1, why would the organism contain alternatives in its genome for the various redox reactions other than the combination yielding maximal ATP/acetate yield? Our analysis of the genome wide network of *C. ljungdahlii* above, has shown that due to the redox enzyme variety and the number of possible ways the redox enzyme can be networked, a combinatorial explosion emerges, leading to no fewer than 24 possible ATP/acetate stoichiometries. After eliminating some of these because they are smaller or equal to zero and others because they are inconsistent with the second law of thermodynamics, 15 feasible gear states hence ATP/acetate production ratios remain. One could argue that the less optimal alternatives are left-overs from a distant past in which life was more difficult. However, over the billions of years of evolution these unused genes would not have any selective advantage and have obtained many mutations rendering them inactive. Perhaps they are inactive, encoding pseudo enzymes [23]. An alternative is that these genes and the proteins they code for still serve some catalytic function yielding a selectable advantage. What could this function be? We posit that these less optimal alternatives might serve as lower speed gears in what could effectively be considered a gear-shifting energy system. Many of the alternatives computed in Table 5.1 should thereto be at the disposition of acetogenic organisms, allowing them smartly to regulate their expression in order to generate appropriate ATP yields, possibly enabling almost seamless gear shifting.

Why would an organism not always go for the most ATP? Gibbs energy dissipation, which goes at the cost of thermodynamic efficiency, is the thermodynamic driver of processes; more than just serving as the arrow of time, flux tends to increase with the increase in Gibbs energy dissipation over a process [24]. Producing more ATP at the same ATP/ADP ratio, coupled to acetogenesis, reduces Gibbs free energy dissipation and as a consequence decreases reaction rates. Thereby living organisms face a rate/efficiency trade-off [1].

To examine this in a simplified setting, we asked the genome wide metabolic map to optimize ATP output flux while producing acetate from CO_2 and hydrogen gas for a range of values for ΔG_{ATP}, the Gibbs energy required in ATP synthesis, representing a changing ATP/ADP ratio, see Figure 5.7. We then recalculated the ATP synthesis flux as a decreasing function of ΔG_{ATP}. By step-wise increases in ΔG_{ATP} the gear-shifting phenomenon can be enforced and shows a sequence of optimal pathways among those highlighted above (Table 5.1) that favor a lower ATP/acetate ratio as the ΔG_{ATP} increases. Although we in a sense artificially construct this sequence of pathways, it shows that if the organism were to maximize a different objective (more on one side of the trade-off) different

pathways (gears) could pop up.

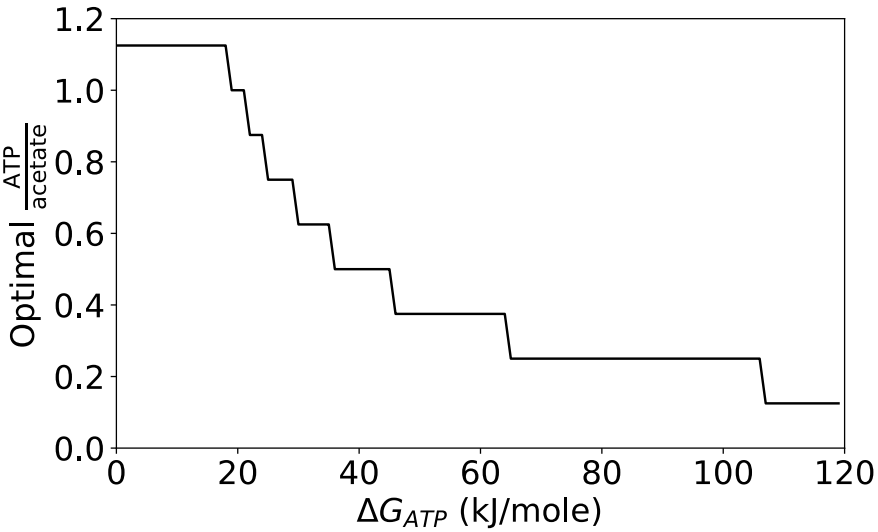


Figure 5.7: Proposed gear shifting for *C. ljungdahlii*: the optimal ratio between ATP and acetate production fluxes for various Gibbs energies of ATP hydrolysis. A genome wide metabolic map for *C. ljungdahlii* [16] was extended with redox reactions identified by Schuchmann and Müller [15] and for a fixed lower bound of carbon dioxide (2) and hydrogen influx (4), requiring the acetate flux to be ≥ 1 (which set that flux at 1), the flux pattern was optimized for ATP output flux. This was done for each combination of electron donors and electron acceptors for hydrogenase, formate dehydrogenase, MTHFR and MTHFD, as present in the extended genome wide metabolic map (see Table 5.1), thus fixing the redox route for each calculation. Calling the ratio of ATP synthesis flux to acetate flux n, the ATP synthesis flux was modelled as $J_{ATP} = n \cdot (1 - n \cdot (\Delta G_{ATP} / \Delta G_A))$ with ΔG_A representing the Gibbs energy released in acetate synthesis from CO_2 and hydrogen ($2CO_2 + 4H_2 \longrightarrow CH_3COOH + 2H_2O$) of -40 kJ/mol taken from Schuchmann and Müller [15]. For each ΔG_{ATP} the n with the highest J_{ATP} was selected as the optimum n and plotted as the ordinate with ΔG_{ATP} as the abscissa.

This degree of dynamics in intracellular bioenergetics through the shuffling of writer domains between enzymes, remains speculative however: it is unclear whether indeed, enzymes can re-associate different writer domains dynamically at a time scale of less than a cell cycle time. Or alternatively, whether they may do this more statically, i.e. during synthesis of the enzyme complexes make their composition depend on post-translational regulation or on relative abundance and transcriptional regulation. Less extensive examples of such gear shifting may have arisen later in evolution where different cytochrome oxidases with different H^+/e^- stoichiometries absorb electrons from the bc1 complex of the cytochrome containing electron transfer chain [25], leading to different growth yields.

Gear-shifting in *S. sulfataricus*

There are other cases in biology where such gear shifting, or at least the possibility to use low gear becomes important under some conditions. An example is again a case where an organism has to deal with extreme conditions, now those of high temperature, the pathway being the lower half of glycolysis. At moderate temperatures the step from glyceraldehyde 3-phosphate to 1,3-bisphosphoglycerate is already problematic in terms of the standard Gibbs free energy difference. This problem is solved by *S. cerevisiae* and by the human, by a proficient subsequent reaction maintaining the concentration of 1,3-bisphoshoglycerate low and thereby the Gibbs energy change across the GAPDH reaction negative. However, at $80\,^{\circ}\mathrm{C}$ this fails to work. By using a dynamic computational model [26, 27] we found that the hyperthermophylic *Archaea S. solfataricus* then, at least in silico, chooses a lower gear (in terms of the number of ATP molecules made per pyruvate produced) by enlisting the non-phosphorylating GAPN reaction for the flux.

We did this as follows: we varied the ATP/ADP ratio by modulating the uni-molecular rate constant of ATP hydrolysis and calculated the fluxes through the GAPDH + PGK route (the pathway with an ATP/pyruvate ratio of 2) and through the GAPN (the non-phosphorylating glyceraldehyde 3-phosphate dehydrogenase which this organism hosts as well) route (the pathway with an ATP/pyruvate ratio of 1). Fig. 5.8 shows that the fraction of the flux between GAP and pyruvate that runs through the low stoichiometry pathway increases with increasing ATP/ADP ratio, hence with increasing free energy of ATP hydrolysis: *in silico* the organism shifts to lower gear when the thermodynamics gets tougher.

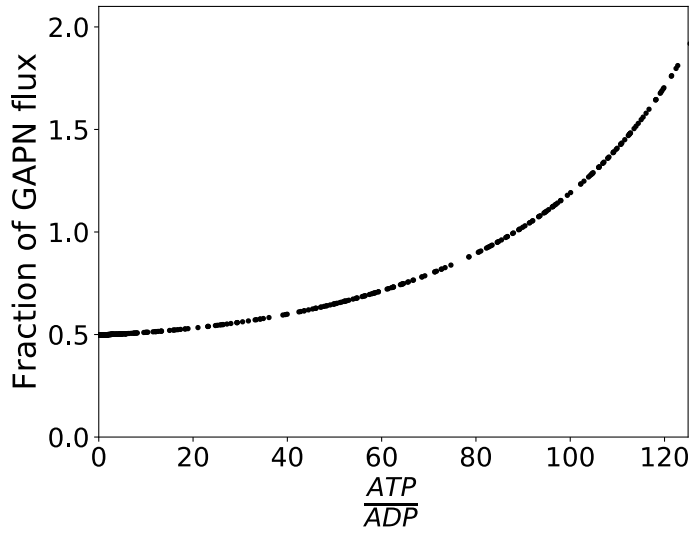


Figure 5.8: The flux fraction from GAP to pyruvate running through the GAPN pathway rather than the GAPDH + PGK pathway, for the *in silico* model of [27]. GAPDH = glyceraldehyde 3 phosphate dehydrogenase. PGK = phosphoglycerate kinase. GAP = glyceraldehyde 3 phosphate. The GAPN activity was set to 0.165 mM/min, the rate constant of the first order ATP hydrolysis reaction was modulated and the ATP/ADP ratio and steady state fluxes calculated.

5.3 Discussion

In this chapter we have revisited non-equilibrium thermodynamics, now in view of metabolic networks that enable various alternative routes and an accompanying variation of stoichiometry at which a desired commodity (such as ATP, biomass or a metabolic product) is produced. We found that variation of the stoichiometry, akin to gear shifting, should enable an organism to produce its commodities at higher rates in changing environments. For one particular condition of anabolic free energy, an organism may have set its network routing so as to achieve the optimal stoichiometry. When conditions are varying such that also the anabolic free energy is affected, e.g. in case of nitrogen starvation setting in, metabolic rerouting such that the stoichiometry shifts, may be advantageous. If such variation in conditions occurs frequently, a continuously varying stoichiometry in accordance with the variomatic principle developed here, might be best: We showed that there should be an optimal mode of rerouting flux through the metabolic network, corresponding to varying the stoichiometry continuously, such that anabolic flux should always be maximal. We call this the 'variomatic' or gear-shifting mode.

We then studied the potential for such stoichiometry variation in the acetogen *C. ljungdahlii*. In this text we have shown that the two propositions by Schuchmann and Müller [15], i.e. that early life on this planet, if exemplified by the acetogen *C. ljungdahlii*, depends on the Wood-Ljungdahl pathway and on electron bifurcation through the Nfn complex, require ramifications if judged from the genome-wide knowledge as associated with the present genome-wide model [16]: The ATP synthesis required can be facilitated by the electron bifurcation in the Nfn complex, but the same result could also be attained by other reactions in the absence of Nfn activity or with reverse flow through Nfn. In some of these cases electron bifurcation then occurs in other reactions such as the methylene-THF reductase. The strict requirement of the Wood-Ljungdahl pathway can also do with some ramification: it is actually the network consisting of the Wood-Ljungdahl pathway, a hydrogenase, the Rnf complex and the H^+-ATPase that is required, where it seems that the formate dehydrogenase can actually be missed, when pyruvate formate lyase digs in (but then additional enzymes are needed as well in order to synthesize the pyruvate from acetylCoA using reduced ferredoxin as electron donor to drive the CO_2 fixation). The ATP yield per acetate through the WLP predicted in [15] can be reproduced using the genome-wide metabolic map of *C. ljungdahlii*, but only after many of the redox enzymes have been bestowed with the substrate and product specificities assumed by Schuchmann and Müller. Furthermore, one should be very precise about proton stoichiometries.

A key assumption in the flux balance analysis portion of this chapter is that we forced the flux through the exchange reaction of acetate to be equal to 1 flux unit in the outward direction. Note that given the medium composition this does force the organism to make full use of both CO_2 and H_2 in the medium: i.e. this confines the GeMM to make acetate from CO_2 and H_2 only and fully. The results presented here would likely change if other compounds would be allowed to be

taken up or secreted.

The main take-away from this is that indeed acetogenic bacteria such as *C. ljungdahlii* (and Archaea such as *S. solfataricus*) may be capable of life at the edge of what is energetically possible by appropriating the optimal gear settings and gear boxes (in terms of electron bifurcation) of the redox network. Therefore, the work shown here may help highlight differences between the current genome-wide reconstructions' predictions and those made by precise microbial biochemistry, and show how differences may be resolved. Finally, the analysis here (and in [15]) suggest multiple interesting targets for strain-design: singularly expressing only those enzyme variants that generate the highest yield of biomass, or of desired metabolic products (Table 5.2).

Pseudo-enzymes and pseudo-signalers may be tips of the iceberg of an evolution that moved forward through shuffling of networks into optimality rather than by evolving proteins to their individual optimality. The reader-writer concept mentioned here and discussed in detail in [19] may help understand the corresponding plasticity, which may also explain the phenomenon of pseudo-enzymes and pseudo-signalers. It should be worth our while to examine further, whether this concept applies to the emerging plethora of pseudoenzymes [28]. The focus on interactions between functionalities that has been introduced by systems biology may improve our understanding of molecular biology, where the word biology is then related to function and fitness.

5.4 Methods

Flux balance analysis

For many simulations in this chapter and the next two chapters, we apply a computational technique called flux balance analysis (FBA) [17]. Briefly (more details are given in the later chapters), this technique concerns the following linear programming problem:

$$\begin{aligned} &\text{maximize or minimize } Y = c^T v, \text{such that for all } k \\ &Sv = 0 \\ &\alpha_k \leq v_k \leq \beta_k. \end{aligned}$$

where S is the stoichiometric matrix for the metabolites, v is the column vector of fluxes through all reactions including exchange reactions with the environment of the system considered, c is a column vector of weights generating the linear combination of fluxes that make up the objective function Y and α and β are the vectors of lower and upper bounds on these fluxes. Superscript T refers to the matrix transpose. A flux distribution returned by FBA is therefore such that each metabolite is produced and consumed or exported in equal amounts, the flux boundaries are accommodated and the flux distribution maximizes (or minimizes) a linear combination of fluxes in the model.

Extending the *C. ljungdahlii* GeMM by Nagarajan et al.

Our starting point was the previously published genome-scale metabolic reconstruction of *Clostridium ljungdahlii* [16]. We obtained the SBML file of the reconstruction from the BiGG database [29] through `http://bigg.ucsd.edu/models/iHN637` on August 25th 2016. This reconstruction covers 698 metabolites, in 785 reactions, encoded by 637 genes. The map had been shown to produce experimentally measured growth rates on various media compositions [16].

In this work we are concerned with reproducing and understanding in more detail the analysis provided by [15] for *C. ljungdahlii* with its genome-wide metabolic reconstruction (abbreviated GeMM for Genome-scale Metabolic Map). To that end we made sure all reactions considered in that study also came to exist in a (slightly) enhanced version of the GeMM, adding reactions where necessary, and we set the simulation conditions appropriately for the acetogenesis problem.

We adjusted the *in-silico* medium to be a mixture of CO_2 and H_2 in a 2:4 ratio by setting the lower bounds on the CO_2 and H_2 exchange reactions to -2 and -4 respectively, where the negative direction indicates uptake of the metabolites. We set the ATP maintenance reaction (ATP $\longrightarrow$ ADP + phosphate) as the objective function with a lower bound of zero, thereby asking with which flux distribution the network could make the ATP synthesis reaction as high as possible. Additionally, we forced the flux through the exchange reaction of acetate to be equal to 1 flux unit in the outward direction. Note that this does force the organism to make full use of both CO_2 and H_2 in the medium: i.e. this confines the GeMM to make acetate from CO_2 and H_2 only and fully. If simulation conditions deviate from those described here, we explicitly highlight the new conditions.

Several reactions described in [15] were added to the metabolic reconstruction to enable the simulations in the scenarios we considered. In Table 5.3 we listed the reactions that were manually added to the reconstruction but which are slightly different in terms of protons from those considered in [15]. In traditional biochemistry considering a single compartment, it is of no importance to keep clear track of protons across reactions since the pH-buffer of the medium is large enough to assuage any problems. When chemiosmotic coupling plays a role however, one needs to keep track of protons that move across the membrane and because membrane potential is often the more important component of the proton motive force other charged species that move across that membrane should also be taken into account. However, because flux-balance analysis requires all species, i.e. also the protons, to be balanced, even one wrongly annotated proton can lead to problems including inaccurate bioenergetics/ATP synthesis. The modeler must account for each of these protons i.e. perform accurate bookkeeping of protons and do so while taking into account the protonation of the metabolites already existing in the reconstruction. Alternatively, one should become explicit in transmembrane charge movement, which is not customary in existing flux balance analysis.

In the reconstruction downloaded from the BiGG database, the ACACT1r (Acetyl-CoA C-acetyltransferase) and HACD1 (3-hydroxyacyl-CoA dehydrogenase) reactions had been blocked, i.e. the lower and upper bound had been set

Reaction ID	Formula
FDHH2	$CO_2 + H_2 \rightarrow Formate^- + H^+$
FDHFDNADPH	$2\,CO_2 + Fd^{2-} + NADPH + H^+ \rightarrow Fd + 2\,Formate^- + NADP^+$
MTHFD_alt	$Methenyl\text{-}THF^{2-} + NADH + Methylene\text{-}THF^{3-} + NAD^+$
MTHFR5_alt	$Methylene\text{-}THF^{3-} + NADH + 2\,H^+ \rightarrow Methyl\text{-}THF^{2-} + NAD^+$

Table 5.3: Reactions added to the genome scale reconstruction from [16].

to zero, for unknown reasons. We unblocked these to allow flux into the beta-hydroxybutyrate synthesis pathway. Finally, we added a demand reaction for removing (S)-3-hydroxybutanoyl-CoA (recycling the CoA) from the cell, so that we may predict its maximal production flux for various network perturbations.

Model checking and visualization

In the main text we perform FBA simulations with various combinations of enzyme alternatives. Ensuring our added reactions did not introduce errors in the energetics, we checked in all considered scenarios of Table 5.1 and 5.2 that the model could not generate Gibbs free energy from nothing. This was accomplished by preventing all medium components from being taken up and then asking for flux through the ATP maintenance reaction. In all cases this returned a maximal yield of zero, meaning the model is not capable of generating Gibbs energy from nothing.

Using the Escher [30] we produced a static map of a subset of the reactions encoded in our (updated) genome-scale metabolic model of *C. ljungdahlii*. This static map can be used, with help of the COBRApy module [31], to visualize flux distributions for any number of situations the modeler may wish to explore. Hence it becomes a fluid map that may generate new visualizations on the fly. Various such images featuring in this work are provided as supplementary files in SVG format. We hope this map may be utilized and extended by others to aid in future work making use of the ljungdahlii genome-wide metabolic map.

Reproducibility

All changes to the published reconstruction were performed through a Python script that loads the original model, performs the changes with aid of the COBRApy package [31] and exports the updated model in SBML version 3 with the FBC package [32, 33]. The Python script, the original and modified model are available as supplementary files. Furthermore, we provide, as a supplementary Github repository, Jupyter notebooks with Python code that reproduces all the analyses discussed in this work. All the discussed models, model analysis code and the visualizations are available on a publicly available Github repositories at: `https://github.com/ThierryMondeel/FOSBE_2016/`, `https://github.com/ThierryMondeel/BST_2017_Gear-shifting` and `https://github.com/ThierryMondeel/FOSBE_2018/`.

References

[1] H. V. Westerhoff and K. van Dam. *Thermodynamics and control of biological free-energy transduction*. Amsterdam: Elsevier, 1987. 10.15490/fairdomhub.1.datafile.4954.1.

[2] G. Nicolis and I. Prigogine. *Self Organization in Nonlinear Systems: From Dissipative Structures to Order Through Fluctuations*. New York: Wiley, New York, 1977.

[3] J. Keizer. *Statistical Thermodynamics of Nonequilibrium Processes*. New York, NY: Springer New York, 1987. 10.1007/978-1-4612-1054-2.

[4] L. Onsager. "Reciprocal Relations in Irreversible Processes. II." *Physical Review* 38 (1931), pp. 2265–2279. 10.1103/PhysRev.38.2265.

[5] L. Onsager. "Reciprocal Relations in Irreversible Processes. I." *Physical Review* 37 (1931), pp. 405–426. 10.1103/PhysRev.37.405.

[6] O. Kedem and S. R. Caplan. "Degree of coupling and its relation to efficiency of energy conversion". *Transactions of the Faraday Society* 61 (1965), p. 1897. 10.1039/tf9656101897.

[7] J. W. Stucki. "The Optimal Efficiency and the Economic Degrees of Coupling of Oxidative Phosphorylation". *European Journal of Biochemistry* 109 (1980), pp. 269–283. 10.1111/j.1432-1033.1980.tb04792.x.

[8] H. V. Westerhoff and Y. D. Chen. "How do enzyme activities control metabolite concentrations? An additional theorem in the theory of metabolic control." *European journal of biochemistry* 142 (1984), pp. 425–30. 10.1111/j.1432-1033.1984.tb08304.x.

[9] B. N. Kholodenko, O. V. Demin, and H. V. Westerhoff. "Control Analysis of Periodic Phenomena in Biological Systems". *The Journal of Physical Chemistry B* 101 (1997), pp. 2070–2081. 10.1021/jp962336u.

[10] W. F. Martin, F. L. Sousa, and N. Lane. "Energy at life's origin". *Science* 344 (2014), pp. 1092–1093. 10.1126/science.1251653.

[11] N. T. Arndt and E. G. Nisbet. "Processes on the Young Earth and the Habitats of Early Life". *Annual Review of Earth and Planetary Sciences* 40 (2012), pp. 521–549. 10.1146/annurev-earth-042711-105316.

[12] H. L. Drake, A. S. Gößner, and S. L. Daniel. "Old acetogens, new light". *Annals of the New York Academy of Sciences*. Vol. 1125. 1. 2008, pp. 100–128. 10.1196/annals.1419.016.

[13] K. Schuchmann and V. Müller. "Energetics and Application of Heterotrophy in Acetogenic Bacteria". *Applied and Environmental Microbiology* 82 (2016). Ed. by A. J. M. Stams, pp. 4056–4069. 10.1128/AEM.00882-16.

[14] K. Schuchmann, J. Vonck, and V. Müller. "A bacterial hydrogen-dependent CO 2 reductase forms filamentous structures". *The FEBS Journal* 283 (2016), pp. 1311–1322. 10.1111/febs.13670.

[15] K. Schuchmann and V. Müller. "Autotrophy at the thermodynamic limit of life: a model for energy conservation in acetogenic bacteria". *Nature Reviews Microbiology* 12 (2014), pp. 809–821. 10.1038/nrmicro3365.

[16] H. Nagarajan *et al.* "Characterizing acetogenic metabolism using a genome-scale metabolic reconstruction of Clostridium ljungdahlii." *Microbial cell factories* 12 (2013), p. 118. 10.1186/1475-2859-12-118.

[17] J. D. Orth, I. Thiele, and B. Ø. Palsson. "What is flux balance analysis?" *Nature Biotechnology* 28 (2010), pp. 245–248. 10.1038/nbt.1614.

[18] W. Buckel and R. K. Thauer. "Energy conservation via electron bifurcating ferredoxin reduction and proton/Na+ translocating ferredoxin oxidation". *Biochimica et Biophysica Acta (BBA) - Bioenergetics* 1827 (2013), pp. 94–113. 10.1016/j.bbabio.2012.07.002.

[19] A. Abudukelimu, T. D. Mondeel, M. Barberis, and H. V. Westerhoff. "Learning to read and write in evolution: from static pseudoenzymes and pseudosignalers to dynamic gear shifters". *Biochemical Society Transactions* 45 (2017), pp. 635–652. 10.1042/BST20160281.

[20] D. Trifunović, K. Schuchmann, and V. Müller. "Ethylene Glycol Metabolism in the Acetogen Acetobacterium woodii". *Journal of Bacteriology* 198 (2016). Ed. by W. W. Metcalf, pp. 1058–1065. 10.1128/JB.00942-15.

[21] F. Baymann *et al.* "The redox protein construction kit: pre-last universal common ancestor evolution of energy-conserving enzymes". *Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences* 358 (2003). Ed. by J. F. Allen and J. A. Raven, pp. 267–274. 10.1098/rstb.2002.1184.

[22] J. M. Murphy, H. Farhan, and P. A. Eyers. "Bio-Zombie: the rise of pseudoenzymes in biology". *Biochemical Society Transactions* 45 (2017), pp. 537–544. 10.1042/BST20160400.

[23] M. Leslie. "'Dead' Enzymes Show Signs of Life". *Science* 340 (2013), pp. 25–27. 10.1126/science.340.6128.25.

[24] R. van der Meer, H. Westerhoff, and K. Van Dam. "Linear relation between rate and thermodynamic force in enzyme-catalyzed reactions". *Biochimica et Biophysica Acta (BBA) - Bioenergetics* 591 (1980), pp. 488–493. 10.1016/0005-2728(80)90179-6.

[25] M. F. Otten *et al.* "Regulation of expression of terminal oxidases in Paracoccus denitrificans". *European Journal of Biochemistry* 268 (2001), pp. 2486–2497. 10.1046/j.1432-1327.2001.02131.x.

[26] T. Kouril *et al.* "Sulfolobus Systems Biology: Cool hot design for metabolic pathways". *Systems Microbiology: Current Topics and Applications* (2012), p. 151.

[27] Y. Zhang *et al.* "The peculiar glycolytic pathway in hyperthermophylic archaea: Understanding its whims by experimentation in silico". *International Journal of Molecular Sciences* 18 (2017). 10.3390/ijms18040876.

[28] V. Reiterer, P. A. Eyers, and H. Farhan. "Day of the dead: pseudokinases and pseudophosphatases in physiology and disease". *Trends in Cell Biology* 24 (2014), pp. 489–505. 10.1016/j.tcb.2014.03.008.

[29] Z. A. King *et al.* "BiGG Models: A platform for integrating, standardizing and sharing genome-scale models". *Nucleic Acids Research* 44 (2016), pp. D515–D522. 10.1093/nar/gkv1049.

[30] Z. A. King *et al.* "Escher: A Web Application for Building, Sharing, and Embedding Data-Rich Visualizations of Biological Pathways". *PLOS Computational Biology* 11 (2015). Ed. by P. P. Gardner, e1004321. 10.1371/journal.pcbi.1004321.

[31] A. Ebrahim, J. A. Lerman, B. O. Palsson, and D. R. Hyduke. "COBRApy: COnstraints-Based Reconstruction and Analysis for Python". *BMC Systems Biology* 7 (2013), p. 74. 10.1186/1752-0509-7-74.

[32] B. G. Olivier and F. T. Bergmann. "The Systems Biology Markup Language (SBML) Level 3 Package: Flux Balance Constraints." *Journal of integrative bioinformatics* 12 (2015), p. 269. 10.2390/biecoll-jib-2015-269.

[33] C. Chaouiya *et al.* "The Systems Biology Markup Language (SBML) Level 3 Package: Qualitative Models, Version 1, Release 1." *Journal of integrative bioinformatics* 12 (2015), p. 270. 10.2390/biecoll-jib-2015-270.

CHAPTER 6

Flux balance analysis for biomarker prediction: A limited proof and in silico test for glutathione mediated drug-detoxification

6.1 Introduction **176**
6.2 Methods **180**
Biomarker prediction with flux variability analysis for IEMs . . 180
Biomarker prediction for drug-metabolism with flux variability analysis 182
Implementation and robustness of the FVA-based approach . . 183
Kinetic biomarker prediction method and simulations 183
Generating the FBA-capable Geenen et al. network 184
Computational reproducibility 184
6.3 Results **184**
Assessing the validity of the BPFVA method 184
Kinetic biomarker predictions 188
The oxoproline loop at the heart of the detoxification pathway . 190
Drug-induced metabolic changes 192
Bypassing the oxoproline loop 192
6.4 Discussion **193**
Supplementary Information **197**

Coauthors:
Vivian Ogundipe, Shintaro Nakayama, Samrina Rehman, and Hans V. Westerhoff

Partially adapted from:
T.D.G.A. Mondeel, V. Ogundipe, H. V. Westerhoff, [Re] Predicting metabolic biomarkers of human inborn errors of metabolism, ReScience. 4 (2018) 1–12. 10.5281/zenodo.1254630

> "Remember that all models are wrong; the practical question is how wrong do they have to be to not be useful."
>
> — George Box [1]

Abstract

Glutathione conjugation in liver is one of the main pathways for the detoxification of reactive metabolites in the liver. Before dosing patients with drugs or other xenobiotics, it should be good to assess the glutathione status of their liver, which is however inaccessible to direct measurements. Instead, metabolites in blood might be used as biomarkers of the liver's glutathione status, but only if they are monotone functions of liver glutathione levels or of capacity. We here test whether a previously proposed method for biomarker prediction that uses flux variability analysis (BPFVA), correctly predicts the suitability of serum ophthalmic acid (OPA) and 5-oxoproline (OXO) as robust biomarkers. We provide the first proof of validity for this method for a subset of possible pathway topologies and then show that the BPFVA method is incorrect in its predictions for a different topology: BPFVA predicts OPA and OXO as biomarkers that decrease in concentration with increasing paracetamol dosage, whilst in both the dynamic computer replica of the network and in experimental datasets, OPA and OXO tended to increase with applied paracetamol, at least in the lower concentration ranges of the latter. The BPFVA method does not reveal that at higher paracetamol concentrations OXO fails as biomarker, and predicts neither the direction of the response of OPA nor that OPA falters as biomarker when the methionine levels are unknown. The results suggest that the BPFVA method is subject to strong limitations, at least for some pathways. We identify a network topology inherent in the glutathione conjugation network that is responsible for the failure of the BPFVA method. The kinetic model of the glutathione pathway, the details and predictions of which are improved here and have been partially validated before, may hold a brighter future in this respect.

6.1 Introduction

The functioning of living cells and organisms is determined by the make-up of their genome, by their immediate environment as well as by the histories thereof. With a genome of roughly 20,000 genes [2], multiple different instantiations of most gene products, and a variety of environmental conditions, this leads to a high-dimensional state space. Although the dimensionality of the total functional space, i.e. the total number of output functions, tends to be much smaller than this, it can still be rather substantial. Due to the extensive networking of most internal processes, the effect of a change in the activity of a gene or in an environmental condition on an output function will depend on the state of all other genes and environmental conditions. It is no wonder therefore that pathologies,

physiologies, as well as responses to medicines differ widely between human individuals [3].

This complexity presents us with a challenge when aiming to predict the effect on health of a genetic variation or change in diet, or when embarking on the inverse problem of network-based identification of drug targets. On the other hand the tools that would seem to enable us to meet this challenge have also become quite impressive. Computation has vastly expanded in terms of speed, storage and accessibility. Genomes can be sequenced, and virtually complete transcriptomes, proteomes and metabolomes may be assessed [4, 5]. The question may therefore be not if, but how one could achieve the predictions aimed for above [6].

The metabolome alone already reveals the dynamics of hundreds of dimensions, probably more than the number of external functions of the network. The metabolome tends to be 'loud' when the fluxome is 'silent' due to redundancies [7]. The metabolome is also pervasive, connecting metabolism and signaling as well as gene expression [8] but not all-encompassing, so that many but perhaps not all genetic or environmental effects on functions should be accompanied by changes in the metabolome.

It is therefore of interest to establish whether and if so, how we are already able to complete tasks such as to (i) predict the effects of a genetic or environmental change on the metabolome, (ii), inversely, infer from an altered metabolome what changes could have caused the alterations, (iii) if one were to know which environmental or genetic change was occurring, infer the extent of the resulting network stress from the measured change in one or a few measurable external metabolite concentrations, which would thereby serve as biomarkers, (iv) predict which changes in environmental conditions could redress a pathological change in network function and (v) predict which internal molecular processes one should affect and how, in order to restore network function (drug therapy design).

Have any modeling methods already been used, or can they be used, to try and accomplish the five tasks mentioned above? The answer is: yes, but always with strong limitations. A 'watchmaker' or 'silicon-cell' dynamic model of the entire intracellular network would be able to simulate the network behavior completely and thereby be able to accomplish all five tasks [9]. Such comprehensive models do not exist although substantial progress has been made, but this has merely resulted in models that were not kinetic at the level of metabolism [10], or in kinetic models of sub-networks only [11, 12]. The underlying limitation is the lack of precisely known rate equations and values for the kinetic constants in them, for the multitude of processes in cell biochemistry. Although requiring fewer parameters, approximate models such as those using power-law rate equations [13] or non-equilibrium thermodynamics-based lin-log equations [14, 15] require fewer parameters, but still more than have been determined experimentally.

What we shall call 'reduced ambition modeling methods' ('RAMM'), have been able to reach conclusions that were incomplete but, for some questions asked, complete enough. One such method, Metabolic Control Analysis (MCA)

has enabled the prediction of the extent to which enzymes control fluxes or concentrations, on the basis of limited information about the components of the network, i.e. the so-called elasticity coefficients, and flux ratios [16]. Through the concentration control coefficients that it computes, MCA should be able to predict which metabolite concentrations increase and which decrease, and even to what extents they should do so, when one reduces the activity of an enzyme in the network [17] (task i) and inversely infer which enzyme activity change had caused an observed change in metabolomics (task ii). Still, for genome scale networks, this would require knowledge of many more elasticity coefficients than is practical.

Because biological functions mostly operate at time scales longer than metabolic transient times, Flux Balance Analysis (FBA) [18, 19] has been able to generate models for the genome-wide analysis of how fluxes might depend on environmental conditions. FBA has been used to predict, successfully, the effect of mutations on growth rate and metabolite production [20]. FBA alone is agnostic of concentrations and describes networks only as fluxes running through them at steady state. Limitations are that those conditions should pertain to fixed input fluxes rather than concentrations and that, still, too many fluxes should be known to be able to solve for the others. The usual approach therefore limits itself to finding the optimal flux patterns with respect to an assumed objective function for constant input fluxes. This predicts maximum yields and the corresponding flux patterns. Even then, for genome wide networks, a great many flux distributions may produce the same maximum yields. Flux Variability Analysis (FVA) [21] then determines the range of every flux in the network that is consistent with attaining that same maximum, allowing other fluxes to compensate. This gives some insight into essentiality and redundancy of specific reactions.

The kinetic modeling methodology has been used for instance to identify limitations to proposed serum biomarkers of glutathione drug detoxification capacity [22, 23]. This was done for a substantial but not yet genome-wide network however and the identification of better combined biomarkers should be treated with the corresponding caution.

Shlomi et al. have developed a way in which FVA [21] is used to predict the effect of mutations on proposed biomarkers [24]. For any network, this approach, which we here call BPFVA (for Biomarker Prediction using FVA), focuses on any 'exchange' reaction in the existing network that connects an extracellular metabolite to a reservoir that we will refer to as the 'medium'. Constraining all network reactions fluxes by upper and lower bounds, the method then determines in silico the range of this exchange's efflux that is compatible with steady state. Shlomi et al. proposed that if this range shifts downward convincingly upon the implementation of a mutation (through inactivation of the corresponding chemical reaction in the network), the concentration of the corresponding metabolite is predicted to decrease with the mutation; the concentration range of the metabolite in the reservoir should shift downward in parallel and the latter concentration may then serve as the biomarker of the effectiveness of the mutation.

Shlomi et al. tested their technique in silico on a kinetic model of red blood cell metabolism [25]. The BPFVA method failed to find most biomarkers: only

40% of the biomarkers evidenced by the kinetic model were identified as such by the BPFVA method ('recall rate'). The biomarkers they did identify were fairly robust however: 73% of the biomarkers predicted by the BPFVA method was also evidenced by the kinetic model ('precision'), i.e. a false positive rate of 27% (100 – 'precision'). Shlomi et al. then applied the BPFVA method to a manually curated set of known inborn-errors of metabolism affecting amino acid metabolism using the OMIM database [26] and found roughly the same false positive rate of biomarker prediction (24%) and an even better (56% 'recall rate') recovery of the experimentally known biomarkers of known inborn errors of metabolism. Thiele et al. [27] found an even better 78% precision using essentially the same BPFVA methodology.

Shlomi et al. did not provide an explicit rationale behind their method nor a proof, not even for a subset of cases. As a consequence, the basis of the method is unclear and thereby it is indeterminate whether the results of the BPFVA method should be expected to depend strongly on particular parameter settings used, such as the on upper and lower bounds of all the reaction rates and whether for some topologies or types of network the method will more successful than for others. In the present study, we shall therefore first examine whether the predictions of the method depend on particular settings of the FBA model, e.g. on the range of admissible velocities of the various reactions (see the Supplementary Information and [28]). In order to provide some more systematic appreciation of when and why the PBFVA method should work, we shall develop a mathematical proof of the validity of the method for a subset of cases.

Extending on this, we shall perform a second in silico validation but now in a network that is home to various mass-balanced cycles, and thereby much different from the erythrocyte metabolic network, i.e. one of the key detoxification pathways for reactive metabolites: glutathione conjugation in the liver. Given that individuals may differ in glutathione levels for various reasons [29, 30] it is crucial not to overdose glutathione-depleting drugs in individuals with low levels of glutathione. It should therefore be important to be able to predict biomarkers of glutathione levels to filter out such individuals before they are treated. These predictive biomarkers could aid in foreseeing an individual's responses to drugs and thereby allow prediction of maximum drug dose levels ultimately assuaging negative consequences.

Biomarkers for the glutathione level have already been predicted using a core kinetic model of glutathione detoxification of acetaminophen (paracetamol) [22, 23]. Here we make use of the first of these published mathematical models [22]. The model represents the essential components of the glutathione conjugation pathway, with connections to proposed biomarkers 5-oxoproline and ophthalmic acid and explicates the interaction with NAPQI (*N*-acetyl-*p*-quinone imine), a toxic by-product produced during the metabolism of paracetamol. In the model by Geenen et al. NAPQI is referred to as 'para' to indicate the source which is considered to be paracetamol. It should be noted that the dose relationship between acetaminophen and NAPQI is not necessarily linear because there are competing metabolic reactions for acetaminophen such as glucuronidation and sulfation and NAPQI may additionally be conjugated by proteins rather

than glutathione.

We revisit the question of whether one should expect 5-oxoproline and ophthalmic acid to be robust biomarkers for glutathione depletion and compare the predictions of the kinetic model with predictions obtained by applying the BPFVA method devised by Shlomi et al. [24] to the same pathway as addressed by the kinetic model.

6.2 Methods

In this work we utilize two biomarker prediction approaches for the same network and intervention therewith: one based on flux balance analysis and flux variability analysis, and one based on kinetic model analysis. Below we will introduce the details of both approaches.

Biomarker prediction with flux variability analysis for IEMs

This section starts by describing the terminology and algorithm of the flux variability analysis (or FVA) approach as originally proposed by Shlomi et al. for finding biomarkers for inborn errors of metabolism [24, 28].

Exchange reactions allow import and/or export of metabolites of the metabolic map. Exchange reactions are represented by non-mass-balanced pseudoreactions, e.g. $X \leftrightarrow \emptyset$. Positive flux indicates net secretion of X by the cell and negative flux indicates the net uptake of X.

We define a boundary metabolite to be a metabolite of which it is known that it may be taken-up by the cell from, and/or secreted to, the medium, which might correspond to the blood or to a reservoir such as the urine. Some boundary metabolites correspond to medium components that are required as input for the cell, e.g. glucose. Others may be produced as waste products, e.g. acetate, whereas yet other species may be subject to either import or export, depending on the specific situation.

Each boundary metabolite will here be associated with an exchange reaction and an exchange interval indicating the enabled range of uptake and secretion fluxes, biochemically set by the total V_{max}'s of all the metabolite's importers and its exporters. This exchange interval is computed through flux variability analysis (FVA) [21], which calculates the range of flux its exchange reaction supports under the following constraints for the flux distributions in the model: (i) network topology, (ii) mass-balance for all internal metabolites, (iii) flux bounds (V_{max}'s) of reactions in the network. Optionally, there is an additional constraint of being optimal with respect to attaining a minimal or maximal flux in certain reactions (e.g. growth), the "objective", in the model.

Mathematically flux variability analysis for a specific exchange reaction v_i

entails the following linear programming problem:

$$\begin{aligned} &\text{maximize or minimize } v_i, \text{ such that for all } k \\ &\mathbf{S} \cdot \mathbf{v} = 0 \\ &\alpha_k \leq v_k \leq \beta_k \\ &Z_{\text{max}} \geq Z \geq \phi \cdot Z_{\text{max}}. \end{aligned}$$

Here $\mathbf{v}$ is the column vector of fluxes representing all reactions in the model including the exchange reactions. v_i is one such an exchange flux of a given boundary metabolite i, α_k is the, possibly negative, lower bound for reaction k and similarly β_k is the, possibly negative, upper bound for reaction k. The reaction bounds are the V_{max}'s of the reactions and are part of the metabolic network definition. Typically they are set to either -1000, 0 or 1000 if the true, biological, V_{max}'s are not known. These values allow specification of irreversible reactions by setting the lower bound (or upper bound) to zero and specification of reversible reactions by setting both the lower and upper bounds to non-zero values. The index i specifies a single reaction that is to be maximized, whereas the index k is used to index the flux bound constraints that exist for each reaction, including reaction v_i. v_i may be maximized in the forward (positive flux) or the reverse (negative flux) direction if allowed by the bounds on v_i. For a network with m metabolites and r reactions, i.e. fluxes, $\mathbf{S}$ is the stoichiometry matrix for the network of size $m \times r$. The numeric, typically integer, elements of $\mathbf{S}$ represent the stoichiometry coefficients of each metabolite m in each reaction r. $Z = \mathbf{c}^T\mathbf{v}$ is the objective function defined for the map, entailing a linear combination, defined by column vector $\mathbf{c}$, of one or more reactions. The vector $\mathbf{c}$ is an indicator of objective reactions, i.e. it contains a value, typically a positive integer (e.g. 1), in rows corresponding to fluxes that are to be included in the objective, and zeros in all other rows. Due to the vector multiplication $\mathbf{c}^T \cdot \mathbf{v}$, the objective Z sums the fluxes of the reactions that correspond to rows containing a value of 1 in $\mathbf{c}$. Z_{max} denotes the maximal value of the objective function. We define $0 \leq \phi \leq 1$ to be an arbitrary minimal fraction of the maximal objective function value that has to be achieved. When $\phi = 1$ we ask for the range of flux allowed for reaction v_i while maintaining the maximal value of the linear combination of objective fluxes. When setting $\phi = 0$, one is asking for the allowable flux range through reaction v_i regardless of the value of the objective function. The latter, i.e. $\phi = 0$, is the case considered in the method by Shlomi et al. and here. The choice of ϕ matters because when $\phi \geq 0$ we are requiring flux to flow through the set of objective reactions (unless $Z_{\text{max}} = 0$). This may entail a forced directionality through reactions that is required to ultimately produce flux in the objective reactions. These additional limitations in the freedom of the flux pattern may subsequently impact the biomarker predictions. The approach discussed here entails predictions purely based on the network topology and therefore uses $\phi = 0$.

Computationally this method entails first calculating the optimal solution value Z_{max} for the objective function Z through flux balance analysis (FBA) and then proceeding to solve the previously described linear programming problem two times for each reaction v_i, once performing a maximization and once per-

forming a minimization of v_i. When ϕ is set to zero the result is independent of the chosen objective fluxes.

This approach allows one to predict exchange flux intervals for each boundary metabolite in a metabolic reconstruction for both standard (wild-type, WT) and variant (e.g. mutant) cases. Shlomi et al. referred to these as the healthy and diseased cases respectively, which was appropriate for the case of genetic mutations leading to inborn-errors of metabolism (IEM). In the results section 6.3 we will at times resort to the general standard (WT) and variant (mutant) terminology with an eye on our drug detoxification application. Shlomi et al. proposed that if any such ranges shift convincingly when the network contains a mutation, the concentration ranges of the corresponding metabolites in the reservoir should shift in parallel and these metabolites would therefore be biomarkers. In other words, biomarkers are those boundary metabolites the range of exchange flux of which differs between the wild-type (WT) and mutant simulations. Shlomi et al. proposed a threshold of at least 10% difference in the lower or upper bounds of the wild-type vs. the mutant flux intervals. Each biomarker will be associated with a prediction for being either elevated or decreased in the mutant case, as compared to the wild type. A biomarker is considered to have an increased, or reduced, extracellular concentration in the mutant case if, when plotting the wild-type and mutant flux-variability intervals on a horizontal axis, both the lower and upper boundary of the mutant interval are shifted to the right, or left, respectively (see Fig. S6.6B and S6.7B). If the two borders of the mutant interval move in opposite directions with respect to the wild-type interval, the result is scored as 'unchanged' and the boundary metabolite is not considered to be a biomarker.

For further details about the implementation of the biomarker prediction algorithm, see the Supplementary Information.

Biomarker prediction for drug-metabolism with flux variability analysis

In the context of inborn errors of metabolism one assumes a specific enzyme to function in the wild-type and to be dysfunctional in the mutant. In the mathematical modelling thereof, one then adjusts the simulation settings to match these conditions [24, 28]. The BPFVA method then computes, for a metabolic reaction r, the exchange flux interval of every boundary metabolite m, both when r is forced to be active (representing the wild-type case), and when r is forced to be inactive (representing the mutant case). However, here we are interested in the network response to influx of a drug compound, which requires a slightly altered approach.

We here extend the BPFVA approach towards drug-metabolism where a drug of interest is taken up by the cell through an exchange reaction. Most details of the approach discussed above for inborn errors of metabolism remain valid. In contrast to the IEM biomarker simulation discussed above, here the standard situation would refer to a case without influx of the drug and the variant situation to a case with influx of the drug. The positions of standard and variant type are thus reversed in terms of presence and absence of a flux. In the drug-metabolism

application of the FVA method, we therefore simulate the WT without forcing any fluxes and by not allowing any drug compound to be taken up by the cell. The mutant simulation is then characterized by forced influx of the drug compound. One of the detoxification strategies in biology involves conjugation of the drug, the conjugate being subsequently exported. As an alternative to drug influx, we might then require the export of the conjugate compound in the mutant simulation.

Implementation and robustness of the FVA-based approach

We implemented the constraint-based biomarker prediction algorithm by Shlomi et al. [24] in a Python function that returns a table with predicted biomarkers. The function has several input parameters most of which are optional but enable the user to perturb the network in various ways and to try different variations of the method, including the drug-metabolism approach. Internally, all flux-variability analyses are performed using COBRApy. [31]

We tested our implementation by reproducing Figures 1, 2 and 3 of Shlomi et al.[24] which entailed an illustrative network and an analysis on biomarker predictions for a selected set of IEMs affecting amino acid metabolism. Our implementation is able to reproduce the original results accurately. We did however, observe that the method is sensitive to various different possible settings of parameters, see Supplementary Information 6.4 and [28].

Kinetic biomarker prediction method and simulations

The kinetic approach to biomarker prediction relies on predicting steady state fluxes of extracellular metabolites towards a reservoir metabolite. One can compare two network configurations one with and one without a specific alteration (e.g. gene deletion, V_{max} reduction or V_{max} activation) and compare the resulting steady state fluxes. Comparison of the two steady state solutions also yields a prediction of the potential change in the biomarker levels. The perturbation we consider is the influx of paracetamol set by the fixed concentration 'para' in the Geenen et al. model.

For kinetic model simulations of glutathione metabolism we used the published glutathione detoxification model [22]. The SBML file was obtained from JWS-ONLINE [32] and imported, run, and analysed in COPASI [33].

Using the steady state results for methionine at $30\,\mu M$ we reproduced figures 2-4 from Geenen et al. (here: Fig. S6.1B,C,D) and for figures 6 and 7 (here: Fig. 6.2C,D) several methionine values were utilized $1\,\mu M$, $15\,\mu M$, $30\,\mu M$, $60\,\mu M$, $100\,\mu M$ and $150\,\mu M$. Additionally, in order to reproduce figures 5 and 8 (here: Figures S6.2A and S6.2B respectively) of Geenen et al., we randomly generated 1000 parameter sets by selecting from uniform distributions of paracetamol in the interval $0\,\mu M$ to $1200\,\mu M$ and methionine ranging from $0.5\,\mu M$ to $100\,\mu M$. For these parameter sets the resulting steady states of the model were calculated. In all cases, the resulting steady state fluxes for ophthalmic acid, 5-oxoproline, GCS, and the intracellular concentration of GSH were recorded and imported into

MATLAB (http://www.mathworks.com/). Using MATLAB, we reproduced the various images that appeared in the original publication. Some differing results were obtained however, see Results section.

Generating the FBA-capable Geenen et al. network

To enable flux variability analysis calculations on the Geenen et al. network, downloaded from JWS online, we had to adjust the network somewhat. This entailed adding paracetamol explicitly as a metabolite (as opposed to a parameter in the kinetic model), and making sure reactions 1, 2, 3, 4 6, 12, 13, 14, 15, 16, 17, 23, 25, 33, 34, 40 and 41 were made irreversible in agreement with the rate laws of the kinetic model. Additionally, we identified a chemically imbalanced reaction 29 where cys = cysgly. We remedied this by adding glycine as an explicit substrate to this reaction. In the kinetic model, *bgly* (blood glycine) is considered to present at a fixed concentration and is taken into account in the rate-law of the reaction but not explicitly in the reaction itself. More such cases exist in this model but then not related to explicitly modeled species.

To aid in understanding the model predictions we used *Escher* [34] to draw a scheme of the network onto which we projected the flux patterns we calculated.

Computational reproducibility

Several Jupyter notebooks [35], MATLAB analysis scripts and COPASI [33] model files used to produce the results and figures in this work are available from the Github repository: https://github.com/ThierryMondeel/Glutathione_biomarkers and will be made accessible through the ISBE infrastructure (see www.ISBE.NL. The repository links to a "MyBinder" project (http://www.mybinder.org) that enables the user to run the Jupyter notebooks without the need to install any software locally, and to reproduce several figures and tables of this paper through an unequivocal procedure. To reproduce the figures made with MATLAB or the analyses performed in COPASI the reader will need the relevant software. The kinetic model from Geenen et al. may additionally be accessed and used from the JWS-ONLINE website (http://jjj.biochem.sun.ac.za/). With this we adhere to desired computational reproducibility practices [36, 37].

6.3 Results

Assessing the validity of the BPFVA method: Theoretical validation for simple linear pathways

Shlomi et al. did not discuss in detail the effects of changing parameter settings of, and assumptions inherent to, their FVA-based method. In the Supplementary Material and [28] we therefore discussed the example of the IEM PKU (phenylketonuria) to highlight that changing parameters and conditions, such as the flux forced through the enzyme in question and the medium definition in the

metabolic map, has potentially large effects on the biomarker predictions. This perspective of biomarkers being potentially less robust than hoped for, calls for methods that readily enable variation of settings such as synchronicity of reaction blocking, the size of the forced flux in the WT, and the cultivation media. Our Python implementation of the algorithm at least partially foresees in these needs.

The algorithm implementation used here expands on this by furthermore allowing biomarker prediction in situations relating to drug metabolism where there is no drug influx into the system in the healthy (WT) conditions whilst in the diseased (mutant) systems there is such an influx (see Methods).

The Shlomi et al. method for predicting biomarkers of reduced activity of metabolic genes has not been proven to be of any kind of general validity. The sole evidence for its at least occasional validity is that it confirms empirically known biomarkers of IEMs [24] as we confirmed in the Supplementary Information. We are here interested in how robust the method is in principle. Perhaps strangely enough the, limited experimental validation of the method is not conclusive then, as it depends on the BPFVA methodology as well as on the completeness of the metabolic map and on the effectiveness of the experimental methodology. It cannot be excluded that failure of both the FVA method and incompleteness of the map would compensate for each other and still lead to correct predictions.

There are only two ways to assess the correctness of the BPFVA method. One is through a proof obtained by analytical mathematics and the second is by examining whether the BPFVA method would successfully predict biomarkers in a completely known system, where the complete knowledge refers to everything needed to describe the relevant variables in the system, i.e. both to carry out the BPFVA and to calculate the behavior of the concentration of proposed biomarkers upon inactivation of any of the reactions. Such systems exist in so-called silicon cells or watchmaker models [9]. These are realistic and complete kinetic descriptions of actual biological systems, which are internally consistent although not necessarily a correct representation of biological reality.

We here start by indeed first providing a novel, but limited, proof of validity of the BPFVA method and then continue by taking the second approach and compare the predictions from the BPFVA approach with those from a kinetic model of glutathione detoxification of paracetamol.

We consider the general case where a metabolite X is synthesized by many reactions 'upstream' and degraded by many 'downstream' reactions, see Fig. 6.1A. We shall also consider the aggregated case (Fig. 6.1B) where all upstream reactions have been taken together in a single reaction and the same for all downstream reactions. We here focus on the case (different from the case studied by Shlomi et al.) where cytosolic X_c is not necessarily transported to the extracellular compartment (denoted by X_e) and is not necessarily exchanged. In cases where in reality there is such an efflux, we use the existing transport and exchange reaction(s) (Fig. 6.1C). Excluding any other reactions, the exchange flux and the transport flux will have the same range of possible values so that we may focus on the transport reaction leading directly out of X_c. When such an

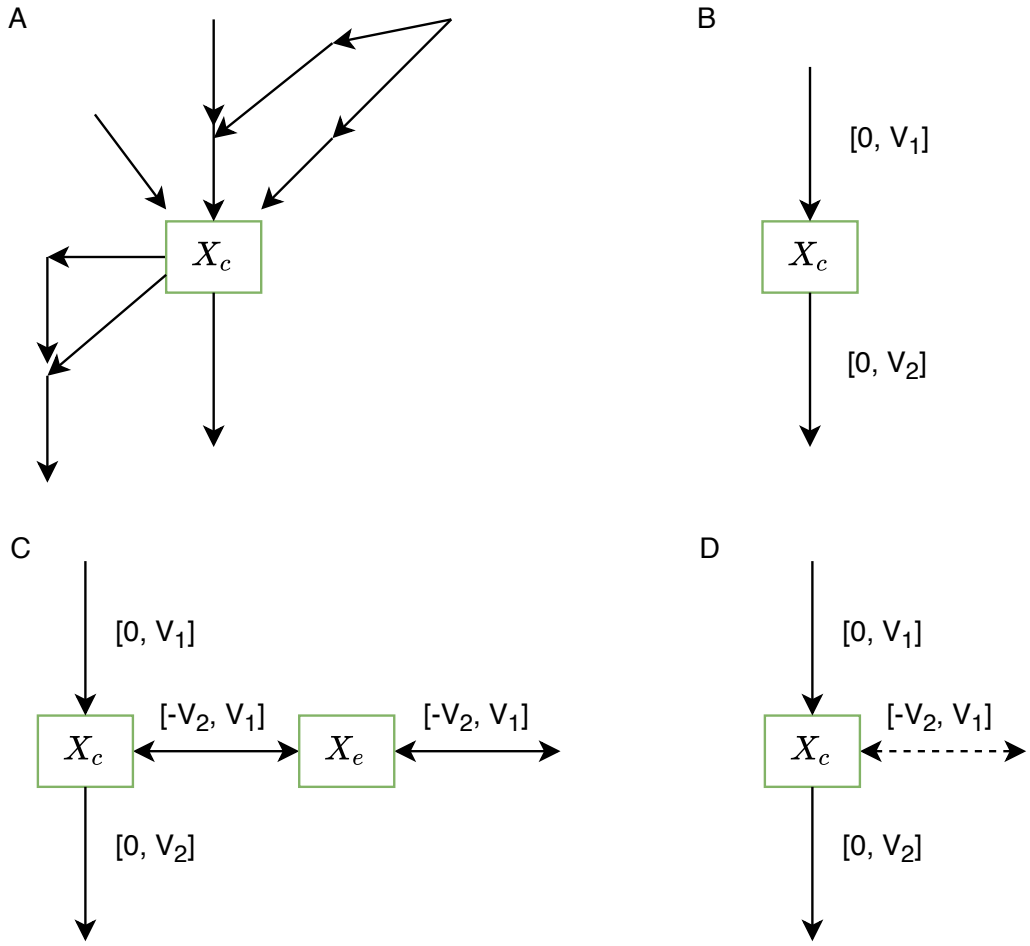


Figure 6.1: Network diagrams for which a rationale for the BPFVA method for predicting biomarkers of activity changes in network enzymes such as in IEMs is here provided. (**A**) Metabolite X exists in a network with various upstream and downstream pathways, the connection between upstream and downstream running through X uniquely. (**B**) The aggregated case where the whole surrounding network has been taken together as a single upstream and a single downstream reaction. The reactions in Fig. 6.1A are delineated such that the concentrations of the substrate of the upper reaction and the product of the lower reaction in Fig. 6.1B are fixed properties. (**C**) When the network contains existing transport and exchange reactions for metabole X we utilize those. X_c refers to the intracellular form of X, X_e refers to the extracellular form of X, which may be exchanged with the environment. In brackets we indicate how the FVA intervals of the transport and exchange reactions depend on the upper bound V_1 of the upstream reaction and the upper bound V_2 of the downstream reaction. (**D**) When no exchange reaction exists for metabolite X we define a virtual exchange reaction (indicated by the dashed line).

efflux is missing in the network, we introduce a virtual exchange reaction for X_c (Fig. 6.1D), which has zero flux at the default steady state. This exchange reaction is not considered real but only used computationally in the FVA. The case consid-

ered here allows us to additionally predict intracellular biomarkers that are not exchanged with e.g. the blood, in contrast to the more limited approach taken by Shlomi et al.

As shown by Westerhoff and Van Dam [15], in terms of metabolic control analysis Fig. 6.1A can be reduced to Fig. 6.1B, if modulations are kept small. Any inactivation of a reaction in the upper set of reactions will therefore correspond to a reduction in the overall afferent flux and hence to a reduction in the aggregated activity, the upper arrow in Fig. 6.1A, and similarly for the efferent reactions. Hence for our purposes here we may focus our attention on Fig. 6.1B.

Focusing on Fig. 6.1B, we assume that the upper reaction has a V_{max} of V_1 and the lower a V_{max} of V_2, whilst both are irreversible. Carrying out a flux variability (Figs. 6.1C and 6.1D) analysis we find that the exchange flux must reside in the interval (-V_2,V_1). Reducing the activity of the upper step by decreasing V_1 will decrease the upper bound, whilst reducing the activity of the lower step will bring the negative lower bound closer to zero.

Inspection of the scheme suggests that inhibiting the upper reaction will decrease the concentration of X, whilst inhibiting the lower reaction will increase X. This then provides an intuitive basis for why a decrease in the upper bound of the upper reaction should be coupled to a decrease in X, making the latter a biomarker of an inhibitory effect on a reaction upstream of X.

We can formalize this intuition as follows: Under the proviso that the elasticity of the lower reaction towards X is positive (i.e. that its rate increases if only the concentration of its substrate is increased, and the elasticity of the upper reaction towards X is negative (i.e. that its rate decreases if the concentration of its product is increased), the control coefficient of the upper reaction on X is positive and the control of the lower reaction negative, for:

$$C^X_{\text{upper}} = -C^X_{\text{lower}} = \frac{1}{\epsilon^2_X - \epsilon^1_X} > 0 \tag{6.1}$$

This result uses the summation and connectivity laws for concentration control coefficients [16] and is strictly valid if there is no flux through the exchange branch (see below). The condition on the elasticities is 'usual' in the sense that substrates usually stimulate reactions and products usually inhibit them. In fact the necessary condition on the elasticities is weaker; merely the difference between the two elasticity coefficients needs to be positive:

$$\epsilon^2_X - \epsilon^1_X > 0. \tag{6.2}$$

This means that there may even be substrate inhibition of the lower reaction provided that the product inhibition of the upper reaction is stronger.

With the above, we deduce that:

$$\text{Decrease in upper FVA bound} \iff \text{Decrease in } V_1 \iff \text{Decrease in X} \tag{6.2b}$$

Hence a decrease in X is a biomarker for inhibition of a step upstream of X. Similarly, an increase in X is a biomarker for inhibition of a step downstream of

X. We are only considering inhibitors here, which may be checked by assessing a reduction in the main pathway flux.

This 'BPFVA method' can even be made quantitative:

$$\frac{dX}{X} = \mathrm{d}\ln(X) = C^X_{\text{upper}} \cdot \mathrm{d}\ln(V_1) = \frac{\mathrm{d}\ln(V_1)}{\epsilon^2_X - \epsilon^1_X}$$

For the simplest case of little product inhibition ($\epsilon^1_X \approx 0$) and approximately linear kinetics (i.e. $v_2 \approx k_2 * X$), this implies that the relative change in the biomarker concentration X is predicted to be equal to the relative change in the upper bound of the FVA, i.e. the change in V_1 brought about by the inhibitor. Should the above approximations not be realistic, then one may still predict the relative change in the biomarker concentration but then requires the magnitudes of the elasticities. For Michaelis-Menten kinetics (i.e. $v_2 = \frac{V_{\max} * X}{K_M + X}$), it holds that $\epsilon^2_X = \frac{K_M}{K_M + X}$ so that when $X = K_M$, the relative change in the biomarker is twice as high as the relative change in FVA bound.

The limitation of the validity of the rule is based in normality (i.e. negativity and positivity, respectively) of the two effective elasticity coefficients (or alternatively, of their difference) and on the scheme of Fig. 6.1A not being completely general.

There is no loss of generality in this proof when considering the transport/exchange reactions as actual rather than "virtual" efflux reactions, which is the approach pioneered by Shlomi et al. without proof. If the flux into the exchange branch is a fraction j of the influx into the system, the two relevant control coefficients are (see supplementary material):

$$C^X_{\text{upper}} = \frac{1}{(1-j)\cdot\epsilon^2_X - \epsilon^1_X + j\cdot\epsilon^3_X}$$
$$C^X_{\text{lower}} = -(1-j)\cdot C^X_{\text{upper}}$$

and are "normally" again positive and negative respectively. Even if j is negative, i.e. there is flux into the system through the exchange branch, this result remains. In the case of linear kinetics, and negligible product inhibition, the two control coefficients are 1 and -1 respectively, independent of the magnitude of the branch flux.

We therefore conclude that predicting biomarkers using the BPFVA method will yield correct results in cases with the above topology and usual elasticities.

Kinetic biomarker predictions for glutathione drug-detoxification

We now turn to our second approach towards assessing the correctness of the FVA method for biomarker prediction, i.e. comparison with a kinetic model, i.e. the model by Geenen et al. To start, we require the kinetic model predictions of how glutathione, oxoproline and ophthalmic acid vary when paracetamol is introduced in the system.

Since these results were already reported by Geenen et al. we worked to reproduce the seven main figures they produced but discovered that some of the

original results were wrongly plotted. In the Supplementary Information we go into detail on the differences between our predictions and the original predictions.

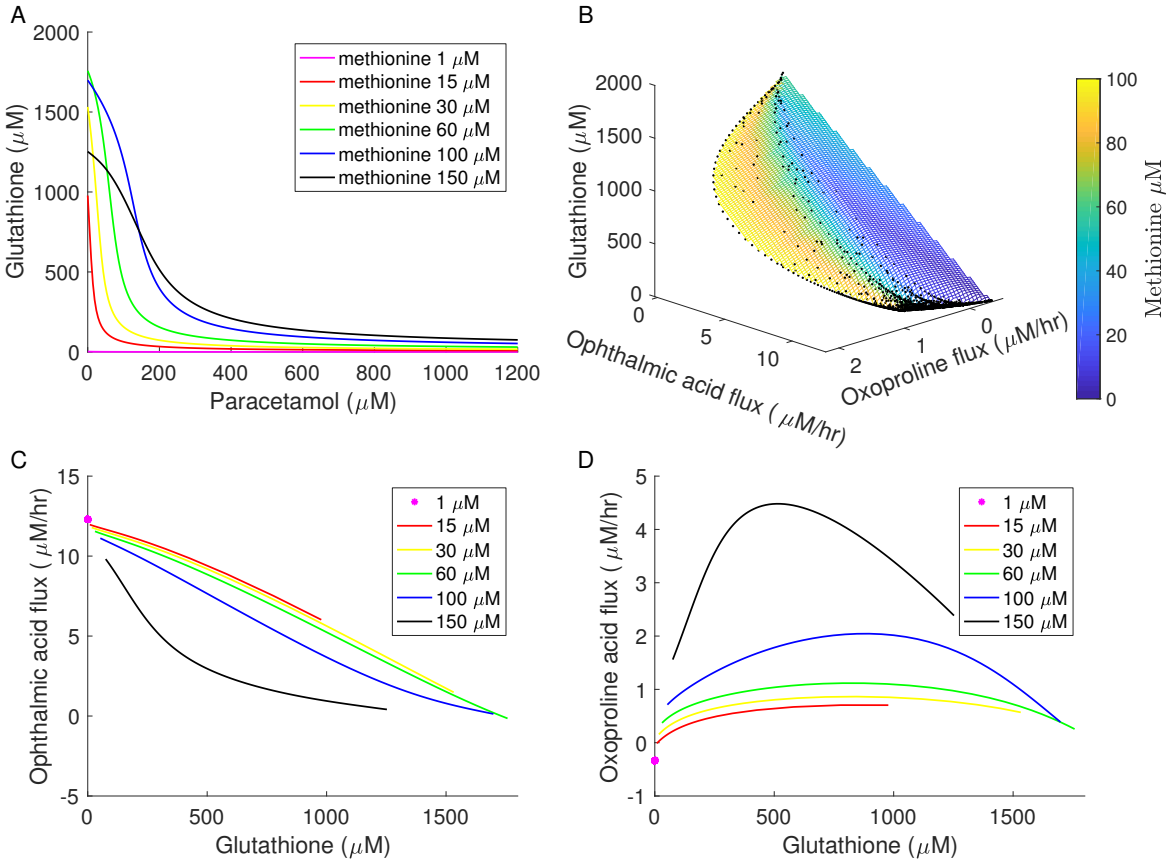


Figure 6.2: Results of computations using the model of Geenen et al. [22]. (**A**) The steady state dependence of glutathione concentration on paracetamol concentration for various concentrations of methionine. (**B**) A 3D plot showing the relationship between ophthalmic acid efflux, 5-oxoproline efflux and steady state intracellular glutathione concentration coloured according to the extracellular methionine level as specified in the colored bar on the right-hand side. (**C + D**) Predicted steady state variation of ophthalmic acid efflux (C) or 5-oxoproline efflux (D) with intracellular glutathione concentration, as paracetamol input was modulated between $0\,\mu M$ to $1200\,\mu M$. In panels A, C and D methionine blood concentration was taken as $1\,\mu M$ (magenta), $15\,\mu M$ (red), $30\,\mu M$ (yellow), $60\,\mu M$ (green), $100\,\mu M$ (blue) and $150\,\mu M$ (black). Note that not all curves address the same range of glutathione concentrations because the parameter being scanned is the paracetamol concentration, for which each value results in a different steady state glutathione level. These glutathione steady state levels are also affected by the methionine concentration and therefore the curves do not span the same domain.

Here, we report on Fig. 6.2 where we jointly display (corrected) reproductions of figures 6, 7 and 8 from Geenen et al. (Figs. 6.2B to 6.2D respectively). To produce Figs. 6.2A, 6.2C and 6.2D we calculated the steady state of the kinetic model for several methionine levels ($1, 15, 30, 60, 100$ and $150\,\mu M$). To produce

Fig. 6.2B (and Fig. S6.2A) we randomly generated 1000 parameter sets by selecting from uniform distributions of paracetamol in the interval $0\,\mu M$ to $1200\,\mu M$ and methionine in the interval $0.5\,\mu M$ to $100\,\mu M$. For these parameter sets the resulting steady states of the model were calculated.

Fig. 6.2A reports the predicted relationship between paracetamol and glutathione for various level of methionine. As can be seen from Fig. 6.2A glutathione steady state levels decrease monotonically with increasing paracetamol for all depicted methionine levels (even for $1\,\mu M$ although it is not clear visually). Futhermore, from Figs. 6.2C and 6.2D we conclude that efflux of ophthalmic acid reduces monotonically with increasing glutathione, in fact for a large range of methionine levels it does so almost linearly, whereas oxoproline varies non-monotonically with increasing glutathione levels depending on the methionine level. The levels of both oxoproline and ophthalmic acid are predicted to depend on methionine levels, but only if the latter are below 0.06 mM. Fig. 6.2B confirms that given the efflux rates of both oxoproline and ophthalmic acid there exists a uniquely predicted intracellular glutathione level as also reported by Geenen et al.

The oxoproline loop at the heart of the detoxification pathway in the Geenen et al. network

Inspired by the reversibility of the 'exchange reactions' in the kinetic model we performed flux balance analysis on a medium that allows uptake of glutamate, glycine, ophthalmic acid, oxoproline and methionine, but not cysteine, all with uptake limits of 10 artificial units (a.u.). We allowed uptake of up to 1000 a.u. of paracetamol and we optimized for the efflux of the cysteine-adduct of paracetamol (i.e. reaction *EX_v37*). We thereby obtained the flux distribution indicated by the orange arrows in Fig. 6.3. This straightforward application of flux balance analysis to the network yielded an unintuitive result: 10 a.u. of paracetamol could be metabolized but only paracetamol and methionine were taken up and an "oxoproline loop" was active utilizing reactions 24 and 31. At steady state in the presence of paracetamol, flux balance analysis predicted that net input of neither glutamate nor glycine is required. This is understandable because the loop recycles the glutamate and glycine contained in the glutathione. Either cysteine or methionine uptake is required since there needs to be a source of the cysteine component in glutathione as this eventually leaves the cell in complex with paracetamol. The computations did not predict cysteine uptake because no such uptake was allowed.

This result is to be juxtaposed with the flux distributions predicted by the kinetic model where the uptake fluxes of glycine and glutamate were active but the oxoproline loop also carried flux. In Fig. S6.3 and Fig. S6.4 we show two steady state flux distributions from the kinetic model both with a blood methionine level of $30\,\mu M$ and for $0\,\mu M$ and $250\,\mu M$ of paracetamol uptake respectively. The 250 µM of paracetamol shifts flux to the detoxification pathway, reduces the steady state level of glutathione (solid curve in Fig. 6.2A) and consequently greatly increased ophthalmic acid production (yellow curve in Fig. 6.2C), and slightly de-

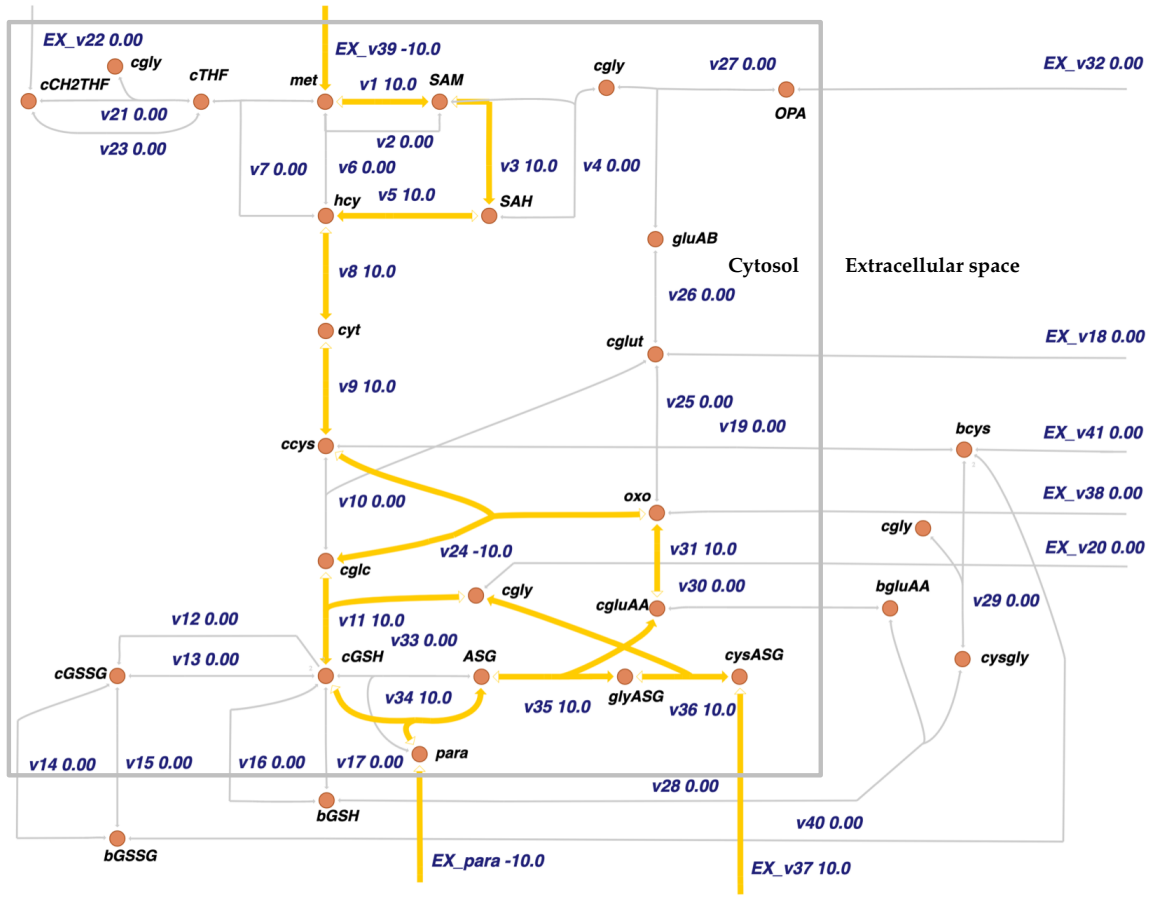


Figure 6.3: Flux distribution, calculated using parsimonious flux balance analysis, of the network from Geenen et al., functioning on a medium with up to 10 a.u. glutamate, glycine, ophthalmic acid, oxoproline and methionine and up to 1000 a.u. of paracetamol. Note the loop-like salvaging of glycine and glutamate from the glutathione-paracetamol complex ()acetaminophen glutathione adduct, abbreviated as ASG), with the remaining cysteine-paracetamol complex (cysASG) being secreted. The fluxes are colored in the spectrum from gray to green to orange where orange indicates high flux, green mean flux (not present in this figure) and gray low flux. Similarly, the arrow thickness increases with the flux magnitude, which is also indicated as a number alongside the arrow. Arrowheads are colored in the direction of the flux (except for EX_para due to technical issues); empty arrowheads refer to the flux running in the opposite direction. Exchange reactions are drawn as if coming from "nothing" in the extracellular compartment. This simply indicates import of the substance into the cell.

creased the 5-oxoproline production (yellow curve in Fig. 6.2D). Additionally, in the kinetic model, the uptake of paracetamol decreases cysteine production because the glutathione is rerouted to paracetamol detoxification.

Drug-induced metabolic changes

Next, we applied the FVA biomarker prediction method using the on/off switch of the paracetamol influx reaction as the modulation point for generating two sets of flux variability intervals.

Importantly, we are not comparing the exact same things between the kinetic and BPFVA methods. In the kinetic case, both Geenen et al. and we looked at biomarkers of steady state cystosolic glutathione levels. Variations in the glutathione levels are obtained by perturbing the network with increasing paracetamol concentrations. In the BPFVA approach we look at biomarkers of the consumption flux of glutathione. Only if glutathione levels vary monotonically with paracetamol influx, the two methods could be the same. As may be inferred from Fig. 6.2A, glutathione decreases monotonically with increasing paracetamol for the whole range of realistic methionine concentrations; this condition is therefore met, at least whenever the system is modulated through varying the concentration of the xenobiotic being detoxified.

To perform the BPFVA we again allowed glutamate, glycine, ophthalmic acid, oxoproline and methionine to be taken up with bounds of 10 a.u. and to be secreted virtually unlimited at an upper bound of 1000. Application of the BPFVA method resulted in the prediction that methionine, methylene-THF (CH2THF in Fig. 6.3) and cysteine are affected (all with reduced extracellular levels) by the increased paracetamol uptake. BPFVA also predicts that glutamate, glycine, oxoproline and ophthalmic acid are not affected. Since methionine (or alternatively cysteine but we did not allow this by setting the lower bound of its exchange reaction to zero) is the only required component to produce glutathione (due to the oxoproline loop) this stands to reason. Methylene-THF shows up because since methionine is required to produce glutathione, less of it may flow to methyl-THF.

But why do 5-oxoproline and ophthalmic acid not show up? An inspection of the network in Fig. 6.3 reveals that both these proposed biomarkers may be readily produced from glutamate and glycine and will therefore only be affected by the addition of paracetamol if these precursors are affected by that addition. The glutamate and glycine salvage loop might therefore be the key reason for this absent prediction prediction of 5-oxoproline and ophthalmic acid as biomarkers. We conclude that for this network topology and for this reason the kinetic and constraint-based approaches of biomarker prediction yield radically different results. This is also supported by the fact that the proof for the validity of the FVA approach given above, does not hold when there are loops and feedbacks in the system that connect subnetworks upstream of the proposed biomarker to subnetworks downstream.

Bypassing the oxoproline loop

Following this result our hypothesis was that if the paracetamol-glutathione (ASG) complex, prior to the salvaging of the glycine and glutamate residues deriving from the glutathione moiety, were exported from the cell this should reveal some effect of increased paracetamol dosage on ophthalmic acid and 5-oxoproline. This situation is more akin to that embodied by the second kinetic

model published by Geenen et al. [23]. We implemented this scenario by adding a reaction to the model that removes the ASG complex and by blocking export of the cysASG complex.

When we applied flux balance analysis to this new situation uptake of glutamate and glycine was mandatory to be able to detoxify paracetamol. The BPFVA method applied to this new scenario revealed that 5-oxoproline and ophthalmic acid should be reduced in the extracellular fluids when paracetamol is added. In addition, glutamate then showed up as biomarker of glutathione levels. These predictions are not in agreement with the kinetic model since in the latter, the 5-oxoproline levels may increase or decrease, depending on the methionine status and on the amount of paracetamol (Fig. 6.2D). The ophthalmic acid prediction is the opposite of the prediction by the kinetic model since Figs. 6.2A and 6.2C of the kinetic model together show that increased paracetamol decreases glutathione levels and therefore should increase ophthalmic acid efflux.

We additionally tested if the flux variability method predicts, like the kinetic model, that the relationship between ophthalmic acid and glutathione depends on the methionine level. We found that increasing the methionine uptake bound in FVA allows more paracetamol to be detoxified, which affects the size of the predicted reduction in ophtalmic acid and oxoproline effluxes but lacked a qualitative effect that flipped the prediction from a reduction to an increase in extracellular levels. Also here the BPFVA fails to predict behavior seen in the kinetic model.

We summarized the results of the predictions of the kinetic and flux-balance-based approaches in Table 6.1.

Biomarker	Kinetic	FVA with efflux as cysASG	FVA with efflux as ASG
5-oxoproline	Increased/Reduced	-	Reduced
Ophthalmic acid	Increased	-	Reduced

Table 6.1: Comparison of biomarker predictions from the kinetic model and the-FVA method for the Geenen et al. network upon increasing the paracetamol uptake. For the kinetic model we deduce the predictions from Fig. 6.3. For the FVA predictions we used the biomarker prediction method comparing exchange intervals before and after adding paracetamol uptake for the case of efflux as ASG and efflux as the cysASG complex. The terms 'increased' and 'reduced' reflect the predicted changes to the extracellular level of the biomarker. For the kinetic model the oxoproline results depend on the methionine and paracetamol levels.

6.4 Discussion

A kinetic model of glutathione metabolism and its detoxification pathway for paracetamol [22] was used here to investigate the reliability of two potential biomarkers, 5-oxoproline and ophthalmic acid, and to investigate the (dis)agreement between a kinetic and an FVA-based biomarker prediction

method for the same network. The work delivered many assets in addition: (i) a Python implementation of the BPFVA method that is applicable to both IEMs and drug metabolism (modeled as a demand for drug excretion as opposed to network mutations) was presented; (ii) insight in the inherent liabilities of the BPFVA method applied to IEMs were discussed in detail and we observed similar issues with the drug metabolism predictions; (iii) a proof of the validity of the BPFVA method was given; (iv) results put forth by Geenen et al. based on a kinetic model of the glutathione metabolism network were revisited, leading to a revision of one of its main conclusions, and (v) a network topology was identified as the culprit of discrepancies between the BPFVA and kinetic model predictions.

In this work we aimed to juxtapose the kinetic model predictions for biomarkers with a method that does not take into account kinetic parameters in any form. To that end we discussed and implemented the constraint-based biomarker prediction methodology proposed by Shlomi et al. and added various options to perturb the network and method. In the Supplementary Information we detailed how we managed to reproduce Shlomi et al.'s results and also illustrated that this method is sensitive to the settings of various of its parameters (e.g. the size of the forced flux, the medium and (a)synchronously blocking affected reactions) and to the network it is applied to. Specifically for the case of phenylketonuria (PKU), the method is sensitive to the details of the model (e.g. Recon 1 vs. Recon 2), the amount of flux forced through the reaction(s) under investigation (or equivalently the setting of the minimal change of the exchange interval) in the wild-type, and the medium settings used in the model. Building on our Python implementation of the BPFVA biomarker prediction method for inborn errors of metabolism we here introduced its extension to drug metabolism. The availability of the implementation we presented here enables other research projects to extend this work.

To predict biomarkers for paracetamol detoxification, we implemented the BPFVA approach to either require the uptake of the drug or, equivalently because there is only 1 detoxification route, the export of the conjugate compound, in the mutant simulation. The latter is effectively a drain of resources on the network resulting in predicted decreased extracellular levels of certain metabolites (in the glutathione case it is a drain on its constituent parts: methionine or cysteine, glutamate and glycine). Note that this is similar to but different from the IEM biomarker prediction case: forcing flux through a typical reaction in a metabolic network (e.g. phenylalaninehydroxylase in PKU) does not (only) force influx of a particular upstream compound (e.g. phenylalanine) but may also lead to forced downstream secretion of another compound (e.g. tyrosine). Lacking the latter property of forced downstream efflux of a compound (other than the conjugate paracetamol complex), the result we found for the BPFVA approach, with all biomarkers that show up being reduced upon increased detoxification, is understandable.

We provide the first public proof of the BPFVA approach correctly predicting qualitative (and even the quantitative) changes in biomarker fluxes, but this proof is only valid for a limited set of network topologies. There may be more topologies for which a proof can be given of the validity of the BPFVA method but we

expect that these proofs are conditional upon elasticity coefficients between the upstream and the downstream modules being small enough or of particular sign. In practice, these elasticities may be unknown, and the BPFVA thereby unreliable.

Of further interest is our finding that the medium defined for the metabolic network, the threshold settings, and the amount of flux forced through the network in the algorithm, all affect the BPFVA outcome significantly (for IEMs in particular). Since the basis for all these settings is feeble, this adds another concern with respect to the robustness of the BPFVA method. These results suggest that the BPFVA is not robust. This emerges both for the case of drug-metabolism modeled as an influx or efflux of a species in the network as introduced here and for the inborn-error of metabolism scenario we dealt with (see the Supplementary Information and [28].

Taking the kinetic model of the glutathione detoxification network for a proof of principle, we found that this network is too complex for the BPFVA method to deliver. This sheds doubt on the BPFVA method. The result however does not address the robustness of the kinetic modeling method. Because the latter comes to subtler conclusions, it may be better in terms of warning for lack of robustness of predicted biomarkers. On the other hand, the kinetic properties that the kinetic method depends on are ill-known for most pathways so that at this stage the kinetic modeling method might be robust in principle, but not usable in practice. The kinetic model considered here did not take into account the distribution of the drug and biomarkers around the body. Such a model does exist and predicted qualitatively similar results as the model considered here [38]. In conclusion, more work is also needed to make the kinetic biomarker prediction method more robust.

The liabilities inherent in the BPFVA method open up the possibility of individualized biomarker predictions. Given an individual- ized metabolic landscape (e.g. by mapping transcriptomics data onto the generic metabolic reconstruction), defining a realistic set of input and output fluxes for a given cell type and individual might enable individualized biomarker predictions, precisely because the method is sensitive to such settings.

An additional outcome of the present work is that some of the (erroneous) results published in the paper Geenen et al. [22] were found to be erroneous and were then revised into corresponding correct results (see Fig. 6.2 and the Supplementary Text): In the original work it was concluded, on the basis of model predictions, that (i) measurement of the secretion flux of either 5-oxoproline or ophthalmic acid in isolation alone is not capable of uniquely identifying the glutathione level within a cell. Moreover, (ii) the relationship between steady-state glutathione levels and the biomarker secretion flux depended on the methionine concentration and (iii) a combined use of both biomarkers resulted in a unique relationship between their secretion fluxes and the glutathione concentration, and that this relationship was independent of the methionine status of the cell. This suggested that only the co-measurement of the two biomarkers should deliver the identifiability of the quantity of interest.

The revision of the Geenen et al. [22] results achieved by the present paper, led to maintenance of conclusions (i) and (iii) (see Fig. 6.2B and 6.2D) but to a

quantitative revision of conclusion (ii): ophthalmic acid continues to be predicted to have a monotonic relationship with glutathione steady state concentration and the relationship can depend on methionine levels (see Fig. 6.2C), but it should not do so for the usual levels of methionine in the blood, which are below 0.03 mM [39]. However, the quantitative character of the modelling results should be dependent on many parameter values that are uncertain. Therefore, quantitative predictions using such a model offer limited certainty. For slightly different parameter values, the dependence on methionine of the covariation of ophthalmic acid flux with glutathione might resurface at lower methionine concentrations than predicted here.

From the updated kinetic analysis we conclude that 5-oxoproline and ophthalmic acid might be candidate biomarkers for glutathione levels. For typical blood levels of methionine, opthalmic acid might be virtually independent of those levels. However, 5-oxoproline does not vary uniquely with glutathione levels and for this reason it is not suitable as biomarker. When measured together 5-oxoproline and ophthalmic acid are predicted to provide a unique predictor of glutathione levels. We do not need to measure methionine in addition to the two biomarkers to predict variations in glutathione levels with varying loads of xenobiotics.

Defining the kinetic model as correct, we found that, and why, the FVA-based method was not robust for this glutathione-based drug detoxification network. There exists, in the known glutathione xenobiotic detoxification network, a cyclic recycling of some of the components of the glutathione-ASG complex. At steady-state this removes the connection of the excretion of this product to 5-oxoproline and ophthalmic acid levels and fluxes. In combination with the steady state condition used in flux balance analysis, this recycling motif (or 'oxoproline loop') leads to unintuitive and false predictions. In a parallel analysis, where the glutathione-ASG complex is secreted prior to such recycling steps (as proposed by Geenen *et al.* in a later model), [23] ophthalmic acid and 5-oxoproline are correctly predicted as biomarkers, however, their qualitative shift is wrongly predicted. This suggests avenues for further research into the possible detection of such network motifs prior to applying the BPFVA method and subsequent improvement of model predictions.

Supplementary Information

On methionine levels

Normal methionine levels in blood are in the range of 0.33 – 0.43 mg/100 ml [39]. At a molecular mass of 149 g/mol, this leads to a blood concentration between 22 and 30 µM. This is still in the range where the methionine concentration hardly influences the ophthalmic acid and 5-oxoproline concentration at any given glutathione level.

Concentration control coefficients in the branched system

Control by enzyme 1 on X

Below, all C's refer to control on X and all ϵ's to elasticity with respect to X. Reaction numbers are with respect to Fig. 6.1D where reaction 1 is the upper reaction, reaction 2 is the lower reaction and reaction 3 the transport or efflux reaction.

We consider a change in the level of enzyme 1 (e_1) (e.g. due to an activator or inhibitor), denoted as $\partial \ln(e_1)$ (the relative change in e_1). This has an effect on the three fluxes J_1, J_2 and J_3 derived below. The steady state flux J is here considered as a function of the three enzyme levels. Partial derivatives taken with respect to one enzyme are thereby taken at constant levels of the other two enzymes, whilst X may vary and the steady state condition is maintained.

First, using the chain rule for logarithmic derivatives

$$\frac{\partial \ln(J_i)}{\partial \ln(e_i)} = \frac{1}{J_i}\frac{\partial J_i}{\partial \ln(e_i)}$$
$$\frac{\partial J_i}{\partial \ln(e_i)} = J_i \cdot \frac{\partial \ln(J_i)}{\partial \ln(e_i)}.$$

Next, using the fact that instantaneous reaction rate v_1 only depends on X, e_1 and parameters p (the substrate is assumed to be a constant) and where the dependence on e_1 is usually a proportionality,

$$\begin{aligned} v_1 &= f(e_1, X(e_1, e_2, \tilde{p}), p) \\ &= e_1 \cdot g(X(e_1, e_2, \tilde{p}), p) \\ \ln(v_1) &= \ln(e_1) + \ln(g(X(e_1, e_2, \tilde{p}), p)). \end{aligned}$$

Consequently when differentiating $\ln(v_1)$ to $\ln(e_1)$ (directly and through X),

$$\begin{aligned} \frac{\partial \ln(J_1)}{\partial \ln(e_1)} &= \frac{\partial \ln(v_1)}{\partial \ln(e_1)} + \frac{\partial \ln(v_1)}{\partial \ln(X)} \cdot \frac{\partial \ln(X)}{\partial \ln(e_1)} \\ &= 1 + \epsilon_X^1 \cdot C_1^X \cdot . \end{aligned}$$

Similarly for v_2 and v_3 (using that $\frac{\partial \ln(e_2)}{\partial \ln(e_1)} = \frac{\partial \ln(e_3)}{\partial \ln(e_1)} = 0$)

$$\frac{\partial \ln(J_2)}{\partial \ln(e_1)} = \epsilon_X^2 \cdot C_1^X$$
$$\frac{\partial \ln(J_3)}{\partial \ln(e_1)} = \epsilon_X^3 \cdot C_1^X.$$

Combining the equations above

$$\frac{\partial J_1}{\partial \ln(e_1)} = J_1 \cdot (1 + \epsilon_X^1 \cdot C_1^X)$$
$$\frac{\partial J_2}{\partial \ln(e_1)} = J_2 \cdot (\epsilon_X^2 \cdot C_1^X)$$
$$\frac{\partial J_3}{\partial \ln(e_1)} = J_3 \cdot (\epsilon_X^3 \cdot C_1^X).$$

Because the (change in) flux into X has to equal the (change in) fluxes out of X

$$\frac{\partial J_1}{\partial \ln(e_1)} = \frac{\partial J_2}{\partial \ln(e_1)} + \frac{\partial J_3}{\partial \ln(e_1)}$$
$$J_1 \cdot (1 + \epsilon_X^1 \cdot C_1^X) = J_2 \cdot (\epsilon_X^2 \cdot C_1^X) + J_3 \cdot (\epsilon_X^3 \cdot C_1^X).$$

Now divide by J_1 to find the expression for the control coefficient of e_1 on X

$$1 + \epsilon_X^1 \cdot C_1^X = (1 - j) \cdot (\epsilon_X^2 \cdot C_1^X) + j \cdot (\epsilon_X^3 \cdot C_1^X)$$
$$1 + \epsilon_X^1 \cdot C_1^X = C_1^X \cdot \left((1 - j) \cdot \epsilon_X^2 + j \cdot \epsilon_X^3\right)$$
$$C_1^X = \frac{1}{-\epsilon_X^1 + (1 - j) \cdot \epsilon_X^2 + j \cdot \epsilon_X^3}$$

where we defined $j = \frac{J_3}{J_1}$ as the fraction of flux going into the efflux branch and we used that since $J_2 = J_1 - J_3$ we have that $\frac{J_2}{J_1} = 1 - \frac{J_3}{J_1} = 1 - j$.

The branching law of concentration control coefficients

Consider a modulation of e_1 and e_3 due to some parameter p with:

$$\frac{\frac{\partial \ln(e_1)}{\partial \ln(p)}}{\frac{\partial \ln(e_3)}{\partial \ln(p)}} = j.$$

where j is the fraction of the flux J_1 flowing into the efflux branch. $\frac{\partial \ln(v_1)}{\partial \ln(e_1)} = 1$, so that for the immediate change in the rates we have:

$$\frac{\partial v_1}{\partial \ln(p)} = J_1 \cdot \frac{\partial \ln(v_1)}{\partial \ln(p)} = J_1 \cdot \frac{\partial \ln(e_1)}{\partial \ln(p)},$$

and

$$\frac{\partial v_3}{\partial \ln(p)} = J_3 \cdot \frac{\partial \ln(e_3)}{\partial \ln(p)} = j \cdot J_1 \cdot \frac{\partial \ln(e_3)}{\partial \ln(p)} = j \cdot J_1 \cdot \frac{1}{j} \cdot \frac{\partial \ln(e_1)}{\partial \ln(p)} = \frac{\partial v_1}{\partial \ln(p)}$$

i.e. the steady state is maintained and therefore $\frac{\partial \ln(X)}{\partial p} = 0$. Hence

$$0 = \frac{\partial \ln(X)}{\partial p} = C_1^X \cdot \frac{\partial \ln(e_1)}{\partial \ln(p)} + C_3^X \cdot \frac{\partial \ln(e_3)}{\partial \ln(p)}$$

Hence the branching law of concentration control coefficients reads:

$$\frac{C_3^X}{C_1^X} = -\frac{\frac{\partial \ln(e_1)}{\partial \ln(p)}}{\frac{\partial \ln(e_3)}{\partial \ln(p)}} = -j$$
$$C_3^X = -j \cdot C_1^X.$$

Additional laws

The summation law [40]:

$$C_1^X + C_2^X + C_3^X = 0$$

The connectivity law [41]:

$$C_1^X \cdot \epsilon_X^1 + C_2^X \cdot \epsilon_X^2 + C_3^X \cdot \epsilon_X^3 = -1$$

Reproducing figures 2-8 from Geenen et al.

We revisited the kinetic analyses previously published in Geenen et al. [22], specifically figures 2-8 from that publication. Using COPASI we scanned the parameter *para* ranging from $0\,\mu M$ to $1200\,\mu M$ in 250 steps. This parameter represents the concentration of paracetamol. We did this for various blood methionine levels ranging from $1\,\mu M$ to $150\,\mu M$. For each combination of these parameters we catalogued the steady state concentrations and fluxes relevant for Figures 2, 3, 4, 6 and 7 of Geenen et al. Using the steady state results for methionine at $30\,\mu M$ we reproduced figures 2-4 from Geenen et al. (here: Figs. S6.1B and S6.1C and part of Fig. S6.1D). For figures 6 and 7 (here: Figs. 6.2C and 6.2D) several methionine values were utilized: $1\,\mu M$, $15\,\mu M$, $30\,\mu M$, $60\,\mu M$, $100\,\mu M$ and $150\,\mu M$.

In order to reproduce figures 5 and 8 of Geenen et al. (here: Figure 6.2B (main text) and Figs. S6.2A and S6.2B), we randomly generated 1000 parameter sets by drawing random numbers from a uniform distribution of paracetamol in the interval $[0\,\mu M - 1200\,\mu M]$ and a uniform distribution of methionine in the interval $[0.5\,\mu M - 100\,\mu M]$. For these parameter sets the resulting steady states of the model were calculated. In all cases, the resulting steady state fluxes for ophthalmic acid, 5-oxoproline, GCS, and the intracellular concentration of GSH were recorded and imported into MATLAB (`http://www.mathworks.com/`). Using MATLAB, we largely reproduced the various images that appeared in the original publication (see below).

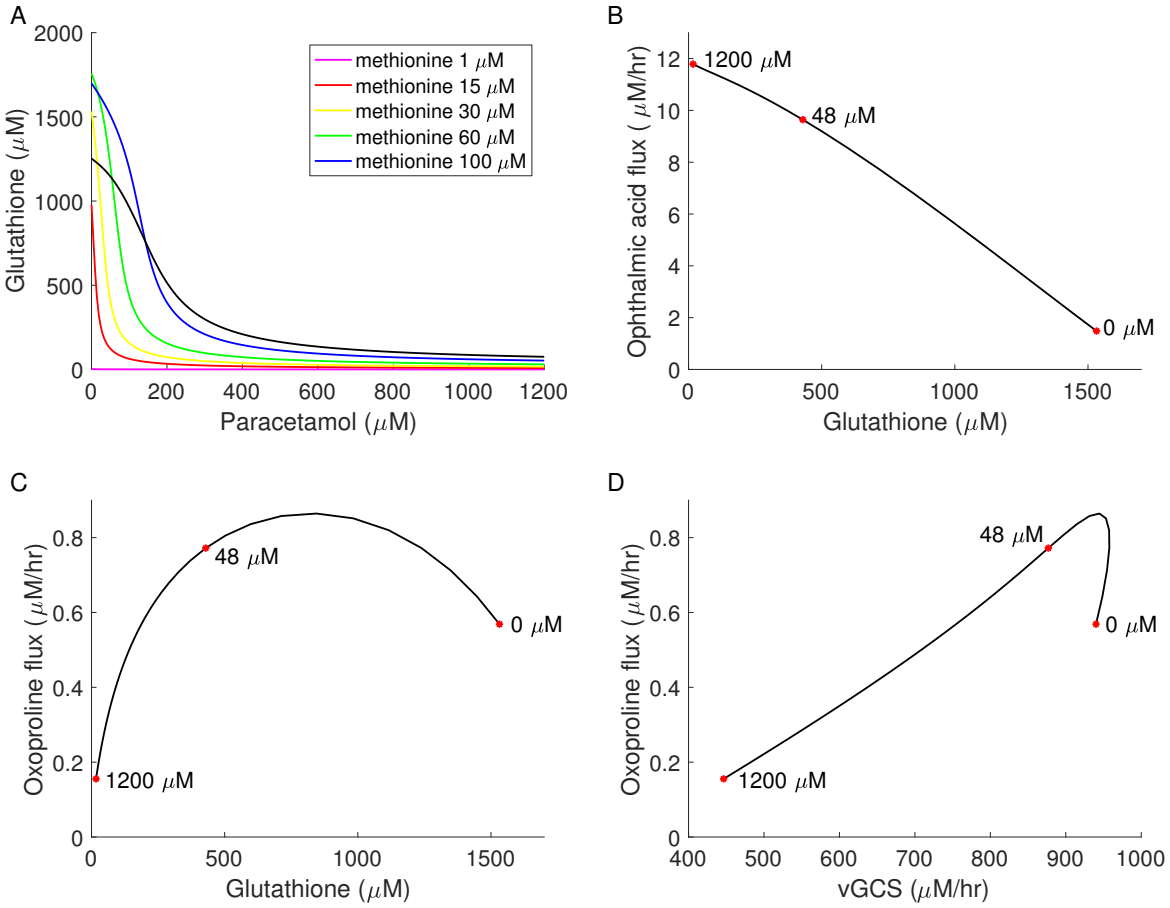


Figure S6.1: Predicted steady state relationships computed for the variation of the concentration of paracetamol. (**A**) The steady state dependence of glutathione concentration on paracetamol concentration for various concentrations of methionine. (**B**) The co-variation of the steady state efflux of ophthalmic acid with intracellular glutathione concentration, for when the paracetamol concentration was varied at $30\,\mu\mathrm{M}$ methionine. (**C**) The co-variation of steady state efflux of 5-oxoproline with intracellular glutathione concentration. (**D**) The co-variation between steady state efflux of 5-oxoproline and the flux through v_{GCS} (v_{10}). Note that in Panels B, C and D v_{GCS} is not the independent variable being varied, paracetamol influx is. We highlight three points in red to indicate the direction of change with regard to paracetamol concentration, which is written next to each point. Panels B, C, and D correspond to figures 2, 3, 4, respectively, of Geenen et al. For (A) the steady states were calculated for a large number of paracetamol concentrations ranging from $0\,\mu\mathrm{M}$ to $1200\,\mu\mathrm{M}$. This was repeated for various methionine concentrations. (B), (C), and (D) steady states were calculated for an external methionine concentration of $30\,\mu\mathrm{M}$. J_X indicates flux of substance X. Molarities are on the basis of intracellular volume.

In the main text and Fig. S6.1A, we highlight a novel figure that clarifies the relationship at steady state between glutathione and paracetamol for various levels of methionine. We observe that glutathione concentration decreases mono-

tonically with increased paracetamol concentration for all methionine levels. For figures 2-4 of Geenen et al., the images we found are similar to their published equivalents and are displayed in Figs. S6.1B to S6.1D. However, we note that in their Figure 4 of the original publication [22] (our Fig. S6.1D) the vGCS-axis (the abscissa) was limited to below $900\,\mu M/h$, which obfuscated interesting behavior past the $900\,\mu M/h$ point where the 5-oxoproline flux suddenly dropped with increasing levels of GCS flux as paracetamol concentration was being reduced. As the indicated paracetamol levels point out, when paracetamol was in the low concentration range (from $0\,\mu M$ to $48\,\mu M$) there is a small effect on GCS flux when paracetamol increases but there is a significant non-monotonic effect on 5-oxoproline efflux. This is the same non-monotonicity as was highlighted by Geenen et al. [22] in their Fig. 3 and our Fig. S6.1.

We failed to exactly reproduce Figures 6 and 7 of Geenen et al., and rather found what we showed in the main text as Fig. 6.2C and 6.2D. Panel 6.2C displays the predicted covariation of ophthalmic acid and glutathione when paracetamol concentrations were increased. In our computations, at zero paracetamol, different methionine levels led to different glutathione levels (Fig. S6.1A) leading to the differing end-point in terms of the abscissa of the curves in Fig. 6.2C. This was not the case in the corresponding Figure 6 of Geenen et al. When decreasing the paracetamol intake and thus increasing glutathione levels (see Fig. S6.1A), we found ophthalmic acid efflux to decrease monotonically as Geenen et al. [22] had found. However, we found that in the methionine interval ranging from $15\,\mu M$ to $60\,\mu M$ it does so approximately linearly in contrast to the original findings by Geenen et al. where a straight decreasing line was only found for the $30\,\mu M$ case. With increasing methionine levels (although for higher levels than in the original) we did recover the non-linear dependencies between ophthalmic acid and glutathione.

Fig. 6.2D concerns the covariation of 5-oxoproline efflux and glutathione levels at steady state when varying the paracetamol concentration. Our recalculation of Figure 7 of Geenen et al. [22] again resulted in different curves as the original with significant quantitative differences with the original. Both these discrepancies (the differences between Fig. 6 and 7 in Geenen et al. and our Fig. 6.2C and 6.2D) could be due to an unfortunate mistake in plotting where abscissa values between the various curves may have been interchanged by Geenen et al. This is a likely explanation, as interrogation of the official version of the model of Geenen et al. [22] (see the legend to its Table 1 and `https://jjj.bio.vu.nl/models/geenen/simulate/`), which was then stored in the live model database JWS-Online, also shows that the level of glutathione at zero paracetamol depends on the level of methionine and does not reach the level of 1500 of the plots in Geenen et al for low blood-methionine levels: for 15 µM of blood methionine ('bmet') the glutathione level at zero paracetamol is computed at JWS-Online as 977, and at 1 µM methionine as 2.1. We tried various combinations of methionine and paracetamol and found the results obtained by JWS-Online and us in the present paper to be the same. In particular we confirmed that at methionine concentrations below 60 µM, the methionine level did not much affect the ophthalmic acid efflux observed at any given glutathione

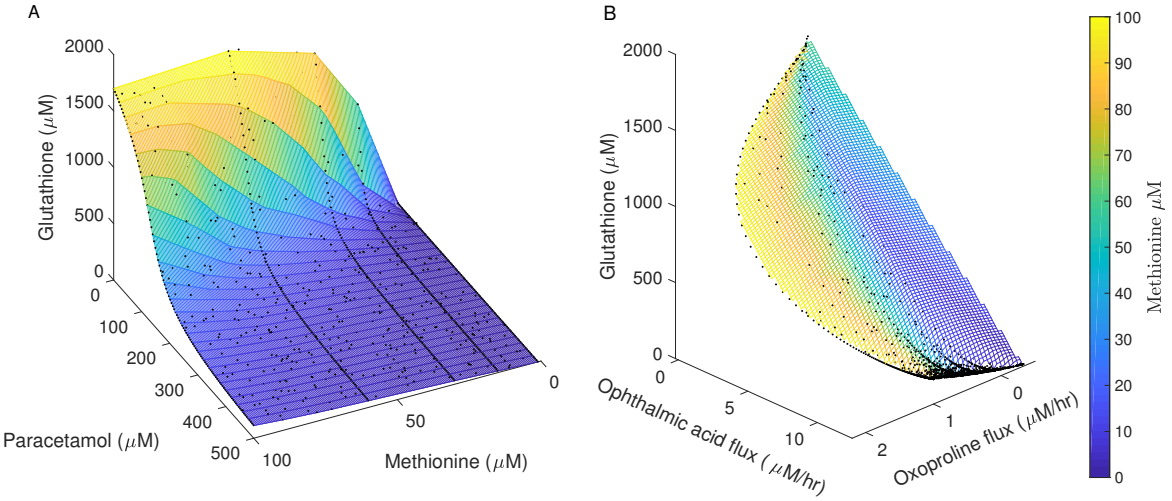


Figure S6.2: (**A**) A 3D plot of the steady state relationship between paracetamol, external methionine and intracellular glutathione. Steady states were calculated by sampling the paracetamol $[0\,\mu\mathrm{M} - 1200\,\mu\mathrm{M}]$ and external methionine concentration $[0.5\,\mu\mathrm{M} - 100\,\mu\mathrm{M}]$ over uniform distributions along their respective intervals. Colouring indicates glutathione level. (**B**) A 3D plot for the same dataset showing the relationship between ophthalmic acid efflux, 5-oxoproline efflux and steady state intracellular glutathione concentration coloured according to the extracellular methionine level.

concentration.

Our results still support the conclusion by Geenen et al. [22] that 5-oxoproline measurements should not be expected to give rise to a unique prediction of glutathione levels, when the methionine concentration is higher than $15\,\mu\mathrm{M}$; at lower methionine concentrations the correlation between oxoproline and glutathione is always positive. In contrast with the conclusions of Geenen et al., measurement of ophthalmic acid levels alone might produce a unique prediction of glutathione levels, independent of the methionine status of the cell: in our hands the model suggested that when methionine is in the range from $15\,\mu\mathrm{M}$ to $60\,\mu\mathrm{M}$ there should be little uncertainty in the glutathione prediction as the curves are very similar. Only for blood methionine levels in excess of $60\,\mu\mathrm{M}$ the predicted glutathione would be uncertain because of dependency on the methionine levels. The usual methionine levels in blood are in the range from $10\,\mu\mathrm{M}$ to $40\,\mu\mathrm{M}$ [39, 42]. Given the uncertainty of kinetic parameters, the original conclusions should still warn against over confidence in ophthalmic acid as sole biomarker for glutathione; for somewhat different parameter values the glutathione-ophthalmic acid covariation might depend on methionine intake.

We reproduced Figures 5 and 8 of Geenen *et al.* as Fig. S6.2. The results are similar to the original. To produce the surfaces shown in Fig. S6.2 we used a linear interpolation of the steady state results for the random samples of paracetamol and methionine concentrations using the *griddata* function in MATLAB. The parameter scan of paracetamol and methionine levels displayed in the right-panel of Fig. S6.2, indicated that the surface made of steady state ophthalmic

acid and 5-oxoproline efflux and glutathione concentration is an unconvoluted straight plane such that two of these quantities may be used to triangulate a single value of the third quantity. Hence, as proposed by Geenen et al., we found that the combined measurement of 5-oxoproline and ophthalmic acid should indicate a unique glutathione state even if the methionine status of the cell were high.

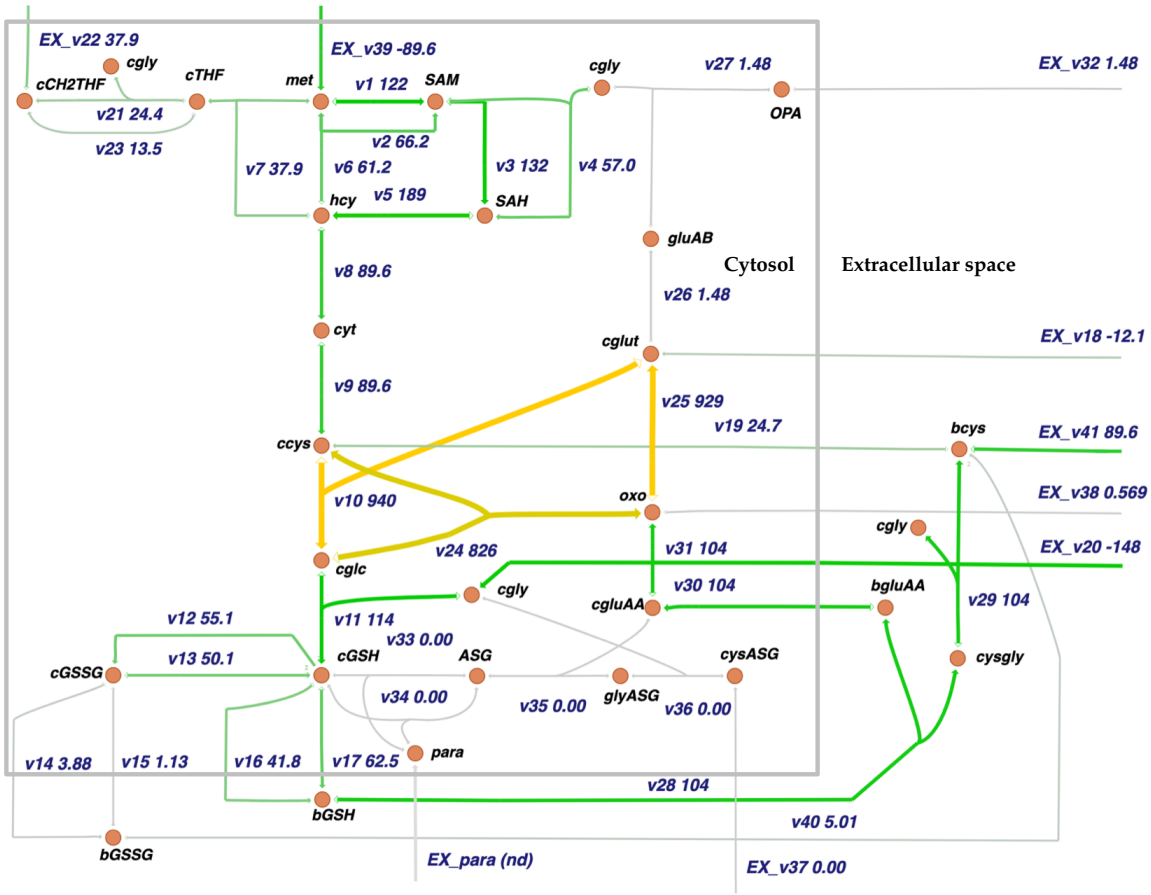


Figure S6.3: Steady state fluxes in the kinetic model by Geenen et al. [22] when paracetamol was absent and blood methionine concentration was $30\,\mu M$ superimposed on the network. The fluxes are colored in the spectrum from gray to green to orange where orange indicates high flux, green mean flux and gray low flux. Similarly, the reaction thickness increases with the flux and flux magnitudes are indicated by the numbers alongside the reacton numbers which are alongside the arrows.

Flux visualizations

In Figs. S6.3 and S6.4 we show fluxes from simulations of the kinetic model from Geenen et al. [22] overlaid on the network for cases of no paracetamol and $250\,\mu M$ paracetamol respectively. In the main text we displayed a similar plot for the fluxes predicted using flux balance analysis.

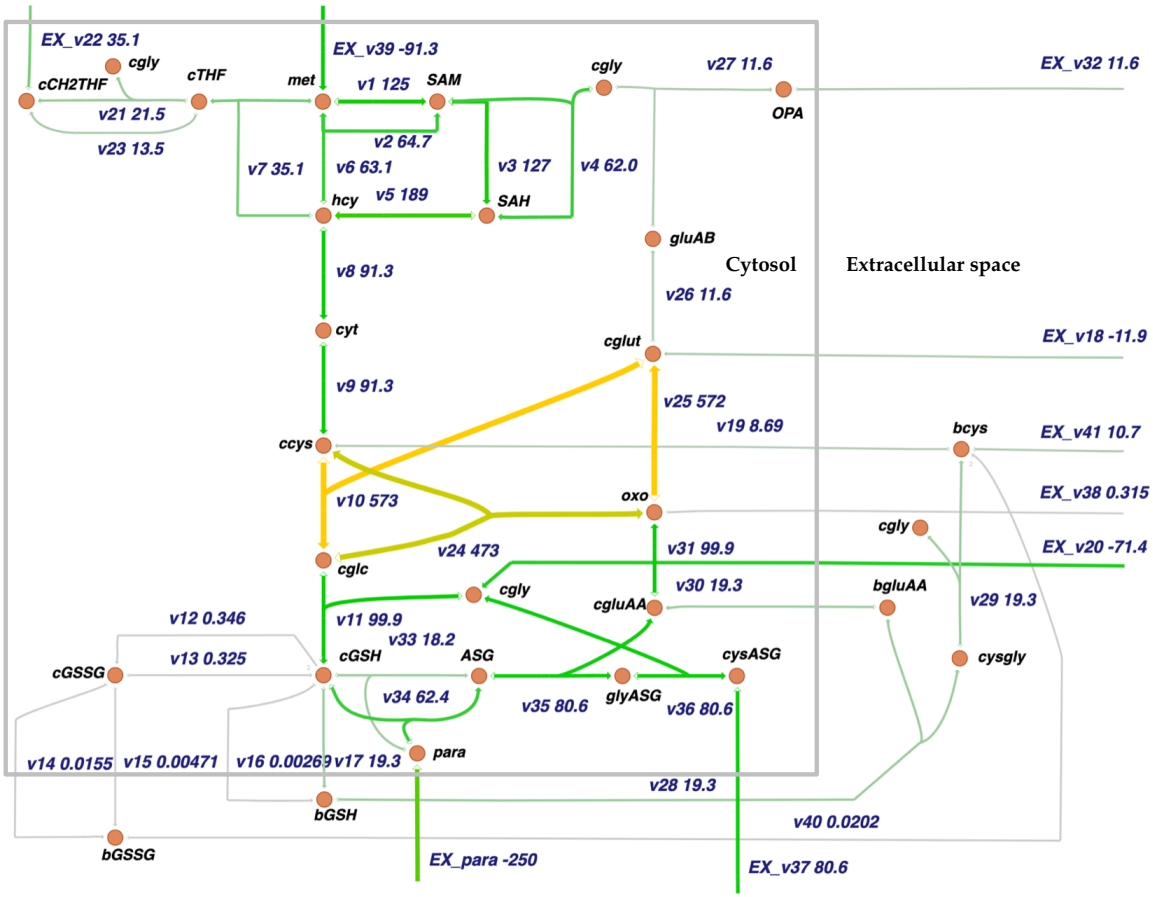


Figure S6.4: Superimposed steady state fluxes on the network for the kinetic model of Geenen et al. [22] when the paracetamol dosage is at $250\,\mu M$ and blood methionine concentration is $30\,\mu M$. The fluxes are colored in the spectrum from gray to green to orange where orange indicates high flux, green mean flux and gray low flux. Similarly, the reaction-arrow thickness increases with the flux.

Reference implementation of Shlomi et al. 2009

The source code for the original publication by Shlomi et al. is not publicly available and the computation environment used was not mentioned in the publication. The MATLAB COBRA toolbox contains code to reproduce the analysis by Thiele et al. but this was also performed in MATLAB.

We here present an implementation of the biomarker prediction method originally proposed by Shlomi et al. , but now programmed in Python, and reproduce the results presented in Figure 1A-B, 2 and 3 of the original publication by Shlomi et al. These figures concern the method's application to a simple metabolic map and then to a set of 17 amino acid metabolism disorders. Our Python module provides the biomarker prediction algorithm and is agnostic of the map used. It can therefore be applied to other maps than the human metabolic reconstruction considered by Shlomi et al. Jupyter notebooks are provided that generate the illustrative map and reproduce all results referred to in Figures 1A-B, 2 and 3 of Shlomi et al. Our implementation is based on information given in the original paper. However, some details critical to the reproduction were left unexplained there and are highlighted here. We also include a more detailed discussion of one of the amino acid IEMs toin which we examine the robustness of the predictions made using this approach. We should add to the usefulness of this reference implementation as well as to the understanding of the approach introduced by Shlomi et al.

We implemented the method of Shlomi et al. in Python. Our Python function has several input parameters, most of which are optional. They enable the user to perturb the network in various ways, which may then differ from the ones implemented here or in Shlomi et al. The default settings are such that they mimic the approach Shlomi et al. originally proposed. All flux-variability analyses were performed using COBRApy [31].

All simulations discussed here were performed in Python 3.6 using COBRApy [31] (version 0.9) and Pandas (version 0.20.3). The accompanying Github repository[1] also functions in conjunction with MyBinder which allows for full reproducibility in the 'cloud' without any need for installation of additional software. The rest of this section introduces the approach proposed by Shlomi et al. as we implemented it.

Here, we examine inborn errors of metabolism (IEMs) resulting from loss-of-function mutations in any single enzyme r that may catalyze multiple chemical reactions. For IEMs that disrupt the activity of several reactions, Shlomi et al. proposed applying the approach to each reaction individually and then combining all the predicted biomarkers into a single list.

Therefore, for a given IEM we run the following sequence of steps for each affected reaction r and for each boundary metabolite m:

1. Compute the exchange flux interval of m, when r is forced to be active with a flux $\epsilon \geq 0$. By default $\epsilon = 1$. We refer to this as the forward wild-type interval $WT^{+}_{r,m}$.

[1] `https://mybinder.org/v2/gh/ThierryMondeel/ReScience-submission-Shlomi2009/Mondeel-Ogundipe-Westerhoff-2017`

2. For reversible reactions r, if the forward interval yields a flux range of $[0, 0]$, constrain the flux to be negative and lower than $-\epsilon \leq 0$, and compute the backward wild-type interval $WT^{-}_{r,m}$.

3. The wild-type interval $WT_{r,m}$ is equal to the union of the forward and backward intervals.

4. Compute the exchange flux interval of m when r is forced to be inactive by temporarily setting the lower and upper bound of r to zero. This interval is denoted by $M_{r,m}$.

5. Compare the wild-type and mutant intervals and predict whether the external concentration of m increases, decreases or remains unchanged when reaction r is forced to be inactive. This is done by the following rule: if both boundaries are higher in the mutant simulation than the corresponding boundaries in the wild-type simulation, then m is predicted to be a biomarker and elevated. If both boundaries are lower in the mutant simulation as compared to the wild-type simulation, then the metabolite is predicted to be a biomarker and reduced. If the predicted changes are smaller than 10%, m is a low confidence biomarker. When the wild-type and mutant intervals are disjoint intervals, m is a highly confident (H.C) biomarker.

After this sequence of steps is completed, and repeated for all r's affected by the mutation or inhibitor, contradictory predictions for m between the affected reactions r are dealt with based on a majority rule. When there is an equal number of elevated and reduced predictions, the boundary metabolite is considered to be unchanged, i.e. not a biomarker.

The flux variability analyses that yield the exchange flux intervals computed in step 1, 2 and 4 are handled by the `flux_analysis.variability.flux_variability_analysis()` function in Cobrapy. The input for this function comprises the model, along with its associated flux bounds which change during step 1, 2 and 4, the list of reactions for which to calculate the flux intervals and the fraction of the optimal objective flux minimally to obtain. In our case we inform the model with flux bounds as described above, lists of exchange reactions for the metabolites of interest, and a fraction of optimum of zero.

Robustness of the BPFVA method

In this section we start with an analysis of the sensitivity of the method which was not clearly discussed in the original publication but is easily facilitated with our reference implementation. Subsequently, we discuss the reproduction of Figures 1, 2 and 3 from the original publication.

Shlomi et al. [24] did not discuss in any detail the effects of changing the parameter settings of the BPFVA approach on the predictions achieved, and they did not present any proof of any generality of the validity of the method. Because our Python implementation allows the user to set various parameter values, it has become easy to examine whether and how parameter settings affect the predictions. An example is given below, where we look at predictions for the

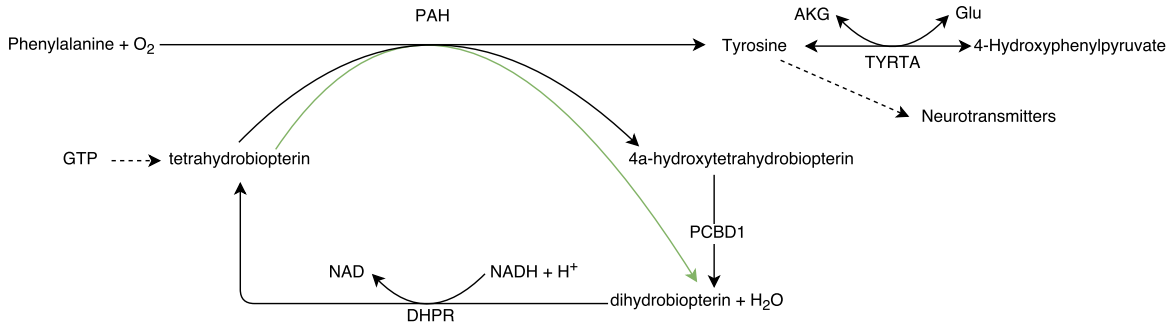


Figure S6.5: Scheme of the network surrounding phenylalanine hydroxylase (PAH) and its cofactors. The 4a-hydroxytetrahydrobiopterin produced by PAH may spontaneously or when catalysed by PCBD1 (EC 4.2.1.96) dehydrate to dihydrobiopterin which may be enzymatically reduced back by NADPH + H^+ to tetrahydrobiopterin through DHPR (EC 1.5.1.34). The solid black lines indicate reactions that are contained in both Recon 1 and Recon 2.2 [43]. The solid green line refers to the alternative PAH reaction that includes the spontaneous dehydradation towards dihydrobiopterin that is only present in Recon 2.2. Dashed lines indicate sources of the synthesis pathways or terminal products of the indicated metabolites. The tyrosine transaminase (TYRTA) uses 2-oxoglutarate = alpha-keto-glutarate (AKG) and glutamate (Glu) as co-substrates for tyrosine and 4-hydroxyphenylpyruvate, respectively.

inborn error of metabolism known as Phenylketonuria (PKU). Here, we simulate this disease by applying one of its potential causes, i.e. dysfunctionality of the enzyme converting phenylalanine into tyrosine.

PKU patients are known for elevated levels of phenylalanine and decreased levels of tyrosine in their serum. We illustrate this briefly in an example also considered in Figure 2 of Shlomi et al. In Fig. S6.5, we sketched some essential features of the metabolic network surrounding the phenylalanine hydroxylase (PAH) enzyme that malfunctions in PKU. Unless stated otherwise, the simulations are performed with a medium that allows efflux through all exchange reactions defined in the metabolic map with an upper bound (for efflux) of 1000 and with a lower bound (for influx) of -1 and we use the default parameter settings, including the threshold of 10% change in the predictions to be labeled a biomarker. The only biomarkers we considered are the 20 proteinogenic amino acids. We considered six cases in total. The results are summarized in Table S6.1.

Perhaps the most fundamental question is: can changes in network topology result in different predictions? The answer is yes. One illustration of this is the comparison of predictions between Recon 1 and Recon 2. As Table S6.1 indicates, using the same medium and algorithm settings for both Recon 1 and Recon 2.2, the two networks led to different predictions of the effect of a defect of phenylalanine hydroxylase on tyrosine. The application of the flux variability method to Recon 1 predicted both tyrosine and phenylalanine as biomarkers as well as the directionality of the change in their concentrations. In contrast, the prediction based on Recon 2.2 fails to identify tyrosine as a biomarker.

Model	R_1	R_2	R_2	R_2 hpp $= 0$	R_2 hpp $= 20$	R_1 $\epsilon = 0$
Synchronicity	(a)sync	async	sync	sync	sync	(a)sync
Phenylalanine	Elevated	Elevated	Elevated	Elevated	Elevated	-
Tyrosine	Reduced	-	Reduced	Reduced*	Reduced$^\wedge$	Reduced

Table S6.1: Summary of predictions made using the BPFVA method for the inborn error of metabolism phenylketonuria for two networks (R_1 = Recon 1 and R_2 = Recon 2), with or without the synchronous setting (indicated by "async" and "sync" or "(a)sync" if the result holds for both) and with high (maximum influx of 20) or low (no influx) hydroxyphenylpyruvate in the medium. * indicates a situation without any tyrosine production capacity in the mutant simulation. $^\wedge$ indicates a situation where tyrosine reduction is so small that it does not reach the 10% threshold. '-' indicates no prediction because the boundaries did not change from wild-type to mutant. $\epsilon = 0$ in the last column refers to the case where we do not force any flux through the affected reaction(s) in the wild-type simulation. In all other columns, $\epsilon = 1$. The full output with interval bounds that the method generated can be found in the relevant notebook in the code repository.

With ever-improving network annotations and consequent metabolic reconstructions, it becomes feasible to simulate the gene knockout directly for most single gene defects leading to inborn errors of metabolism. Shlomi et al. directly blocked or activated reactions associated with a causal gene for any of a number of specific IEMs. For enzymes partaking in several reactions, Shlomi et al. proposed activating and blocking all reactions associated with the gene individually and then taking the union of their predicted biomarkers. A majority rule was applied when a biomarker was subject to qualitatively different predictions between the enzyme-catalyzed reactions encoded by the gene. We will refer to this as the asynchronous setting of BPFVA. The alternative of blocking all reactions affected by the mutation or inhibitor simultaneously could yield different predictions. We will refer to this as the synchronous setting. In our implementation, we added a parameter to choose between the asynchronous and synchronous setting.

For example, consider a gene that encodes an enzyme that catalyzes two chemically different reactions, R_1 and R_2. Using the asynchronous setting, we would run the algorithm once for R_1 and once for R_2 and take the union of all predicted biomarkers. Inconsistent predictions between R_1 and R_2 are dealt with based on a majority rule (see Methods). In contrast, with the synchronous setting, we would run the algorithm once. For the wild-type simulation, we would simultaneously force flux through both reactions and in the mutant simulation, we would block both reactions. The synchronous approach simulates the complete catalytic inactivity or absence of the enzyme and therefore the infeasibility of any reactions it catalyzes. This corresponds to a deletion of the gene. The asynchronous approach might correspond to a point mutation that changes the specificity of the enzyme.

Returning to the example of phenylketonuria, we simulated this for Recon 2.2 using the synchronous setting. Column 3 in Table S6.1 shows that the result

then agrees with the Recon 1 predictions. The results may be understood by the following argumentation: Using Recon 1, we came to the correct prediction because it contains only one reaction ('PHETHPTOX2') linked to the PKU gene, i.e. only phenylalanine hydroxylase (Entrez ID: 5053, HGNC:8582):

$$\begin{gathered}\text{(6R)-L-erythro-5,6,7,8-tetrahydrobiopterin} + \text{L-phenylalanine} + O_2 \\ \updownarrow \\ \text{4a-hydroxy-L-erythro-5,6,7,8-tetrahydrobiopterin} + \text{L-tyrosine}\end{gathered}$$

In Recon 1, the 4a-hydroxytetrahydrobiopterin produced may dehydrate to 6,7-dihydrobiopterin by the 'THBPT4ACAMDASE' reaction catalyzed by the enzyme PCBD1 (EC 4.2.1.96). In addition to the PAH and PCBD1 catalyzed reactions indicated above, Recon 2.2 contains the overall reaction where the spontaneous dehydration reaction is included, again attributed to phenylalanine hydroxylase:

$$\begin{gathered}\text{(6R)-L-erythro-5,6,7,8-tetrahydrobiopterin} + \text{L-phenylalanine} + O_2 \\ \updownarrow \\ \text{6,7-dihydrobiopterin} + \text{L-tyrosine} + H_2O.\end{gathered}$$

These details are included in the pathway summary in our Fig. S6.5. When the reactions affected by the PKU gene are knocked out simultaneously, Recon 2 predicts both tyrosine and phenylalanine as biomarkers. When blocking only one reaction, a way to produce tyrosine remains. Consequently, tyrosine is not predicted as a biomarker if Recon 2 is used as the metabolic map.

Another significant modulation point for the method, and for flux-balance-based models in general, is the medium composition, which represents the nutrition. It matters for the biomarker prediction which other exchange reactions are allowed to have a non-zero flux and what the bounds on these fluxes are: nutrition matters for the effects of mutations on function. Changes in medium composition may open up or shut down entire parts of the metabolic network that rely on specific substrates. Additionally, such changes interact with the significance threshold, since the medium components together with the network pathways (and their V_{max}'s) determine the amount of a given metabolite that can be produced.

In order to examine these ramifications, we continue with the synchronous PKU example and note that tyrosine production flux goes down in the mutant simulation, but not to zero. This is due to an alternative tyrosine synthesis pathway in the model originating from 4-Hydroxyphenylpyruvate in the medium through the 'TYRTA' reaction (tyrosine transaminase E.C. 2.6.1.5). This transaminase is considered to be reversible in Recon 2. 'Equilibrator' [44] annotates this reaction with a $K'_{eq} \approx 1$. It should be mentioned, however, that the ability for transport of hydroxyphenylpyruvate across the cell membrane is speculative. Columns 5 and 6 in Table S6.1 show the results obtained when using the standard medium and a medium without or with increased (i.e. influx at -20) 4-hydroxyphenylpyruvate influx respectively. In FBA and FVA, changes in

medium concentrations of exchange metabolites are modelled by changing the corresponding influx (inward exchange) V_{max} (bound). In the former situation, i.e. without increased 4-hydroxyphenylpyruvate influx the ability to produce tyrosine indeed disappears as a consequence of an inactivation of PAH. In the latter situation, i.e. with increased 4-hydroxyphenylpyruvate influx, there is so much leftover capacity for making tyrosine that the change in interval bounds slips below the 10% cutoff. This sensitivity to medium composition appropriately reflects the phenomenon of cells behaving differently in different culture media, and organisms behaving differently depending on their nutrition. Our discussion here may highlight that this feature should be expected to affect the pertinence of biomarkers. Biomarkers may have a tendency to be non-robust, dependent as they can be on network topology details and nutrition.

Finally, we consider the value of the parameter that sets the amount of flux forced through the reaction of interest in the wild-type condition. The final column in Table S6.1 indicates that phenylalanine does not show up as a biomarker if when setting $\epsilon = 0$ and the PKU simulation is performed on Recon 1. This may also be the reason why Sahoo et al. did not manage to find phenylalanine as a biomarker for PKU in their study [45].

Reproduction of panel A and B of Figure 1 in [24]

The original publication's Figure 1, panel A and B, exemplified the biomarker prediction method for an illustrative network. It is reproduced here as Fig. S6.6. The simple nature of the example serves to understand the rationale of the method and also helped us validate our implementation of the method. This repository provides an SBML file of the network, the Jupyter notebook that generates it and a Jupyter notebook reproducing the biomarker prediction. The image itself was produced using Adobe Illustrator.

The network in Fig. S6.6A consists of seven metabolites, 6 of which (i.e. all except M3) have exchange reactions. We consider a hypothetical deletion of the enzyme catalyzing the conversion of M1 into M2 coupled to the conversion of M3 into M4. Shlomi et al. graphically provide the biomarker prediction for this network in their Figure 1A. Here, we do so graphically and numerically in Fig. S6.6B and this paper is accompanied by a Jupyter notebook reproducing the biomarker prediction. In our analysis, all exchange reactions were given inward bounds of -10 and outward bounds of $+1000$ so that the internal reaction bounds (of ± 1000) are never reached. Fig. S6.6B shows that all exchangeable metabolites except M5 and M7 are predicted to be robust biomarkers.

The exchange interval we found for metabolite M6 shows that in the wild-type case, M6 can be either taken up from biofluids or secreted back into them. The upper bound (at 9) is lower in absolute value than the lower bound (at -10) because in the wild type, at least 1 unit of flux is needed to provide M3 as substrate for the enzyme under investigation since the enzyme converting M1 and M3 to M2 and M4 is required to have a minimal flux of 1 unit. In the disease case, M6 (synthesized through M5) can only be secreted to biofluids at a maximal flux (10) equal to minus the bound of EX_M5. These results are in full agreement

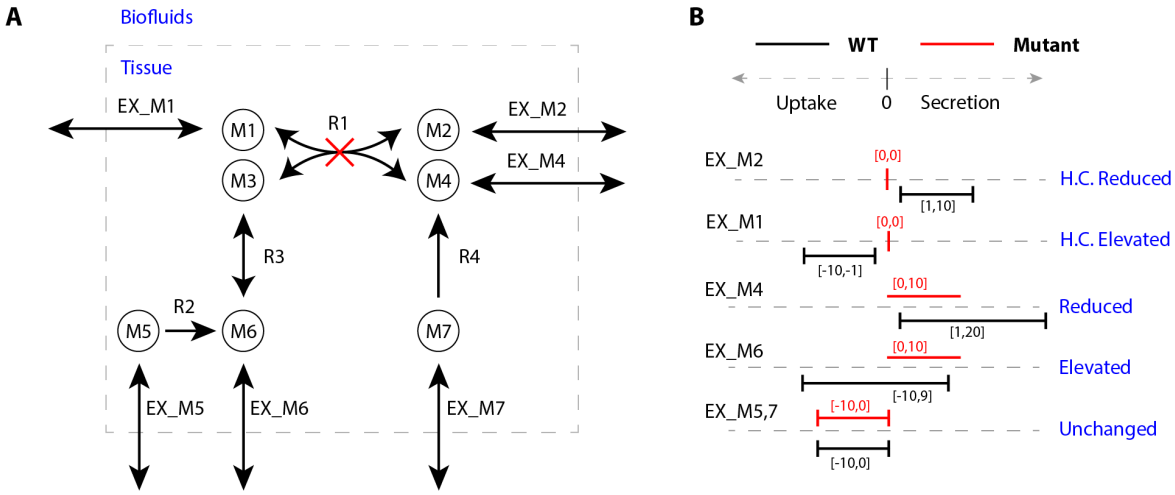


Figure S6.6: (**A**) A metabolic network used to illustrate the predictions of biomarkers of IEMs as practiced by Shlomi et al. [24]. All metabolites but M3 have an exchange reaction allowing for their uptake or secretion. M2 synthesis is fully dependent on uptake of M1 and M5 or M6. The IEM considered is one that leads to deletion of the enzyme converting M1 plus M3 to M2 plus M4. (**B**) Illustration of the predicted exchange intervals for each of the metabolites with an exchange reaction. The wild-type (WT) interval is indicated in black, the mutant interval in red. Each interval is associated with a numeric interval predicted by the BPFVA algorithm (indicated between square brackets) and with a qualitative prediction for the level of the metabolite in the biofluids (indicated in blue on the right; see the Methods section). H.C. is an abbreviation for a highly confident biomarker (see main text).

with the original publication.

Reproduction of Figure 2 in [24]

Shlomi et al. applied their method to the reconstructed human metabolic map, examining a set of 17 IEMs affecting amino-acid metabolism. Figure 2 in their publication compares predicted biomarkers with known biomarkers from the OMIM database [26]. We reproduced the analysis for each IEM listed in Table S6.2 and found full agreement between what was reported by Shlomi et al. and the results of our implementation, but only for a certain setting of the medium composition. Shlomi et al. did not report on their chosen medium in their publication but we were able to reproduce their results by setting all inward flux bounds to -10 and all outward flux bounds to 1000.

We made use of the Excel file included as supplementary information in [24] and included it in this repository as well. The genes and reactions linked to a given IEM are detailed in that Excel file. We attempted to extract the genes for each IEM and using the Recon 1 map, deduced the coupled reactions, but found out that this yielded predictions different from the ones presented by Shlomi et al. However, the reactions listed in the Excel file were not all the same reactions as those that are linked in Recon 1 to the genes linked to the IEMs. Assuming the reconstruction of the metabolic network to be accurate, this raises some doubts as to the correctness of the results obtained by Shlomi et al. for those IEMs where the listed reactions did not agree with the reconstructed map.

Additionally, we needed to add manually the information for S-Adenosylhomocysteine hydrolase and Methionine adenosyltransferase deficiency, since they were missing from the Excel file, and to reduce the tetrahydrobiopterin deficiency gene list so as to only consider QDPR (quinoid dihydropteridine reductase). Lastly, Histidinemia was in the Excel file associated to the reaction HISD whereas this should have referred to HISDr, i.e. to match the reaction identifier in Recon 1. After these changes, we recovered the results as originally presented and as summarized here in Table S6.2.

Reproduction of Figure 3 in [24]

Figure 3 in the original publication entailed an in-depth look at the effect of inborn errors in AHCY or CBS on methionine metabolism. We reproduced this figure as Fig. S6.7. We used Adobe Illustrator to redraw the network from the original publication and based the intervals in panel B on the numerical results the algorithm predicted. The qualitative results are already contained in Table S6.2. The numerical results illustrated in Fig. S6.7B may be reproduced using the notebook *Reproduce_figure_2_and_3_Shlomi2009.ipynb*. For clarity we included a section in the notebook that details the results relevant to Fig. S6.7B.

We note that in early versions of this reproduction we were utilising a medium that allowed influx of -1 for all medium components, as opposed to -10 now. A setting of -1 does accurately reproduce the results in Figure 2, but not in Figure 3. For the CBS IEM the WT interval would correspond to $[-1, -1]$ and would therefore be a point instead of a wide interval as Shlomi et al. draw it.

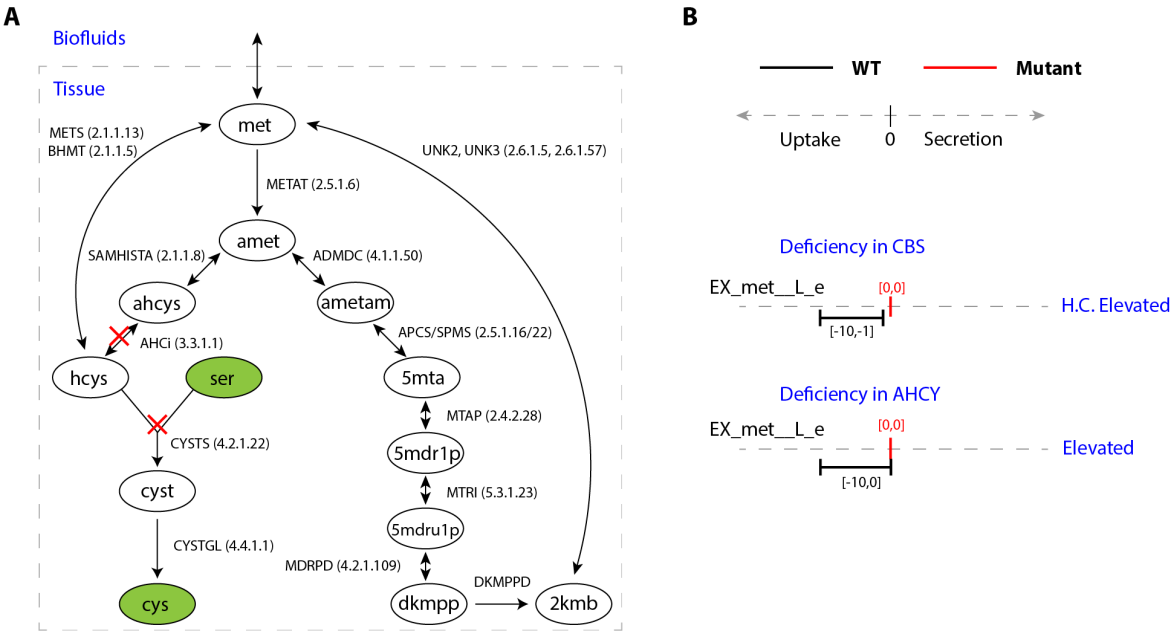


Figure S6.7: Reproduction of Figure 3 in [24]. (**A**) Illustration of the sub-network of Recon 1 relevant to the effect of homocystinuria on the metabolism and transport of methionine. Nodes represent metabolites and edges represent reactions. For simplicity, only abbreviations of metabolite names are given in the figure. The full names are: L-Methionine (met), S-Adenosyl-L-methionine (amet), S-Adenosyl-L-homocysteine (ahcys), L-Homocysteine (hcys), L-Serine (ser), L-Cystathionine (cyst), L-Cysteine (cys), S-Adenosylmethioninamine (ametam), 5-Methylthioadenosine (5mta), 5-Methylthio-5-deoxy-D-ribose 1-phosphate (5mdr1p), 5-Methylthio-5-deoxy-D-ribulose 1-phosphate (dmdru1p), 2,3-diketo-5-methylthio-1-phosphopentane (dkmpp), 2-keto-4-methylthiobutyrate (2kmb). Metabolites with a green background color participate in additional reactions not shown here. Reactions are indicated by their identifiers in Recon 1 as well as by a corresponding enzyme commision number if available. We highlighted the mutations for Homocystinuria (dysfunctional CBS = CYSTS = EC 4.2.1.22) and hypermethioninemia (dysfunctional AHCY = AHCi = E.C. 3.3.1.1). (**B**) Prediction of concentration changes of methionine in homocystinuria or S-adenosylhomocysteine hydrolase deficiency based on the biomarker prediction algorithm. The wild-type (WT) interval is indicated in black, the mutant interval in red, with BPFVA predicted bounds indicated between square brackets and the corresponding qualitative prediction for the level of the metabolite in the biofluids in blue lettering (see Methods). H.C. is an abbreviation for a highly confident biomarker, see main text.

IEM	Affected reactions	Elevated	Reduced
IEM	Affected reactions	Elevated	Reduced
S-Adenosylhomocysteine hydrolase	SEAHCYSHYD, AHCi	L-Methionine	L-Cysteine
Alkaptonuria	HGNTOR	L-Tyrosine	
Argininemia	ARGN		
Cystinuria	CYSTSERex, SERLYSNaex		
Lysinuric protein intolerance	SERLYSNaex		
Glutamate formimino-transferase deficiency	FTCD, GluForTx	L-Histidine	
Histidinemia	HISDr	L-Histidine	
Homocystinuria	CYSTS, MTHFR3, METS	L-Methionine	L-Cysteine
Hyperprolinemia type I	PRO1xm, PROD2m		
Maple syrup urine disease	OIVD1m, OIVD2m, OIVD3m	L-Valine, L-Isoleucine L-Leucine,	
Methionine adenosyl-transferase deficiency	SELMETAT, METAT	L-Methionine	L-Cysteine
methylmalonic acidemia	CBLATm, CBL2tm, MMMm	L-Isoleucine	
Phenylketonuria	PHETHPTOX2	L-Phenylalanine	L-Tyrosine
Phenylketonuria II	DHPR		
Tyrosinemia type I	FUMAC	L-Tyrosine	
Tyrosinemia type III	PPOR, 34HPPOR	L-Tyrosine	
Glycine encephalopathy	GCC2am, GCC2bim, GCC2cm, GCCam, GCCbim, GCCcm		

Table S6.2: Biomarkers predicted by the BPFVA method. We used our implementation to reproduce the biomarkers for the list of inborn errors of metabolism shown in Figure 2 in [24]. Using the supplementary Excel table of [24] linking IEMs to causative genes and affected reactions, we filtered out those that are listed here. The only biomarkers considered were the 20 proteinogenic amino acids. We report all biomarkers the change of which exceeds the 10% threshold and categorize them by their qualitative prediction: elevated or reduced serum levels.

The reason for this stems from the forced flux $\epsilon = 1$ in WT. When the maximum influx of methionine is -1 (as a medium component) it is also required at a flux of -1 due to the forced active flux in the WT. When the maximum influx is increased to -10 the lower bound of the predicted interval increases to -10 but the upper bound does not because ϵ is still equal to 1.

We included this simulation in the last cell of the notebook. This is another example of the inherent sensitivity of the BPFVA approach to various settings in the algorithm.

References

[1] G. E. P. Box and N. R. Draper. *Empirical model-building and response surfaces.* New York: John Wiley & Sons, 1987.

[2] S. L. Salzberg. "Open questions: How many genes do we have?" *BMC Biology* 16 (2018), p. 94. 10.1186/s12915-018-0564-x.

[3] W. E. Evans and M. V. Relling. "Moving towards individualized medicine with pharmacogenomics". *Nature* 429 (2004), pp. 464–468. 10.1038/nature02626.

[4] B. Berger, N. M. Daniels, and Y. W. Yu. "Computational biology in the 21st century". *Communications of the ACM* 59 (2016), pp. 72–80. 10.1145/2957324.

[5] P. J. Denning and T. G. Lewis. "Exponential laws of computing growth". *Communications of the ACM* 60 (2016), pp. 54–65. 10.1145/2976758.

[6] H. V. Westerhoff, S. Nakayama, T. D. Mondeel, and M. Barberis. "Systems Pharmacology: An opinion on how to turn the impossible into grand challenges". *Drug Discovery Today: Technologies* 15 (2015), pp. 23–31. 10.1016/j.ddtec.2015.06.006.

[7] L. M. Raamsdonk *et al.* "A functional genomics strategy that uses metabolome data to reveal the phenotype of silent mutations". *Nature Biotechnology* 19 (2001), pp. 45–50. 10.1038/83496.

[8] B. H. ter Kuile and H. V. Westerhoff. "Transcriptome meets metabolome: hierarchical and metabolic regulation of the glycolytic pathway". *FEBS Letters* 500 (2001), pp. 169–171. 10.1016/S0014-5793(01)02613-8.

[9] J. L. Snoep, F. Bruggeman, B. G. Olivier, and H. V. Westerhoff. "Towards building the silicon cell: A modular approach". *Biosystems* 83 (2006), pp. 207–216. 10.1016/j.biosystems.2005.07.006.

[10] J. R. Karr *et al.* "A Whole-Cell Computational Model Predicts Phenotype from Genotype". *Cell* 150 (2012), pp. 389–401. 10.1016/j.cell.2012.05.044.

[11] B. M. Bakker, P. A. M. Michels, F. R. Opperdoes, and H. V. Westerhoff. "Glycolysis in Bloodstream Form Trypanosoma brucei Can Be Understood in Terms of the Kinetics of the Glycolytic Enzymes". *Journal of Biological Chemistry* 272 (1997), pp. 3207–3215. 10.1074/jbc.272.6.3207.

[12] B. Teusink *et al.* "Can yeast glycolysis be understood in terms of in vitro kinetics of the constituent enzymes? Testing biochemistry". *European Journal of Biochemistry* 267 (2000), pp. 5313–5329. 10.1046/j.1432-1327.2000.01527.x.

[13] M. A. Savageau. "Introduction to S-systems and the underlying power-law formalism". *Mathematical and Computer Modelling* 11 (1988), pp. 546–551. 10.1016/0895-7177(88)90553-5.

[14] D. Visser and J. J. Heijnen. "Dynamic simulation and metabolic re-design of a branched pathway using linlog kinetics". *Metabolic Engineering* 5 (2003), pp. 164–176. 10.1016/S1096-7176(03)00025-9.

[15] H. V. Westerhoff and K. van Dam. *Thermodynamics and control of biological free-energy transduction*. Amsterdam: Elsevier, 1987. 10.15490/fairdomhub.1.datafile.4954.1.

[16] D. B. Kell and H. V. Westerhoff. "Metabolic control theory: its role in microbiology and biotechnology". *FEMS Microbiology Letters* 39 (1986), pp. 305–320. 10.1111/j.1574-6968.1986.tb01863.x.

[17] B. N. Kholodenko, H. V. Westerhoff, J. Schwaber, and M. Cascante. "Engineering a Living Cell to Desired Metabolite Concentrations and Fluxes: Pathways with Multifunctional Enzymes". *Metabolic Engineering* 2 (2000), pp. 1–13. 10.1006/mben.1999.0132.

[18] E. P. Gianchandani, A. K. Chavali, and J. A. Papin. "The application of flux balance analysis in systems biology". *Wiley Interdisciplinary Reviews: Systems Biology and Medicine* 2 (2010), pp. 372–382. 10.1002/wsbm.60.

[19] J. D. Orth, I. Thiele, and B. Ø. Palsson. "What is flux balance analysis?" *Nature Biotechnology* 28 (2010), pp. 245–248. 10.1038/nbt.1614.

[20] A. P. Burgard, P. Pharkya, and C. D. Maranas. "Optknock: A bilevel programming framework for identifying gene knockout strategies for microbial strain optimization". *Biotechnology and Bioengineering* 84 (2003), pp. 647–657. 10.1002/bit.10803.

[21] R. Mahadevan and C. Schilling. "The effects of alternate optimal solutions in constraint-based genome-scale metabolic models". *Metabolic Engineering* 5 (2003), pp. 264–276. 10.1016/j.ymben.2003.09.002.

[22] S. Geenen *et al.* "A mathematical modelling approach to assessing the reliability of biomarkers of glutathione metabolism". *European Journal of Pharmaceutical Sciences* 46 (2012), pp. 233–243. 10.1016/j.ejps.2011.08.017.

[23] S. Geenen *et al.* "Glutathione metabolism modeling: A mechanism for liver drug-robustness and a new biomarker strategy". *Biochimica et Biophysica Acta (BBA) - General Subjects* 1830 (2013), pp. 4943–4959. 10.1016/j.bbagen.2013.04.014.

[24] T. Shlomi, M. N. Cabili, and E. Ruppin. "Predicting metabolic biomarkers of human inborn errors of metabolism". *Molecular Systems Biology* 5 (2009), p. 263. 10.1038/msb.2009.22.

[25] N. Jamshidi *et al.* "Dynamic simulation of the human red blood cell metabolic network". *Bioinformatics* 17 (2001), pp. 286–287. 10.1093/bioinformatics/17.3.286.

[26] V. A. McKusick. "Mendelian Inheritance in Man and Its Online Version, OMIM". *The American Journal of Human Genetics* 80 (2007), pp. 588–604. 10.1086/514346.

[27] I. Thiele *et al.* "A community-driven global reconstruction of human metabolism". *Nature Biotechnology* 31 (2013), pp. 419–425. 10.1038/nbt.2488.

[28] T. D. G. A. Mondeel, V. Ogundipe, and H. V. Westerhoff. "[Re] Predicting metabolic biomarkers of human inborn errors of metabolism". *ReScience* 4 (2018), pp. 1–12. 10.5281/zenodo.1254629.

[29] S. C. Lu. "Regulation of glutathione synthesis". *Molecular Aspects of Medicine* 30 (2009), pp. 42–59. 10.1016/j.mam.2008.05.005.

[30] S. C. Lu. "Glutathione synthesis". *Biochimica et Biophysica Acta (BBA) - General Subjects* 1830 (2013), pp. 3143–3153. 10.1016/j.bbagen.2012.09.008.

[31] A. Ebrahim, J. A. Lerman, B. O. Palsson, and D. R. Hyduke. "COBRApy: COnstraints-Based Reconstruction and Analysis for Python". *BMC Systems Biology* 7 (2013), p. 74. 10.1186/1752-0509-7-74.

[32] B. G. Olivier and J. L. Snoep. "Web-based kinetic modelling using JWS Online". *Bioinformatics* 20 (2004), pp. 2143–2144. 10.1093/bioinformatics/bth200.

[33] S. Hoops *et al.* "COPASI–a COmplex PAthway SImulator". *Bioinformatics* 22 (2006), pp. 3067–3074. 10.1093/bioinformatics/btl485.

[34] Z. A. King *et al.* "Escher: A Web Application for Building, Sharing, and Embedding Data-Rich Visualizations of Biological Pathways". *PLOS Computational Biology* 11 (2015). Ed. by P. P. Gardner, e1004321. 10.1371/journal.pcbi.1004321.

[35] T. Kluyver *et al.* "Jupyter Notebooks—a publishing format for reproducible computational workflows". *Positioning and Power in Academic Publishing: Players, Agents and Agendas* (2016), pp. 87–90. 10.3233/978-1-61499-649-1-87.

[36] R. D. Peng. "Reproducible Research in Computational Science". *Science* 334 (2011), pp. 1226–1227. 10.1126/science.1213847.

[37] G. K. Sandve, A. Nekrutenko, J. Taylor, and E. Hovig. "Ten Simple Rules for Reproducible Computational Research". *PLoS Computational Biology* 9 (2013). Ed. by P. E. Bourne, e1003285. 10.1371/journal.pcbi.1003285.

[38] S. Geenen *et al.* "Multiscale modelling approach combining a kinetic model of glutathione metabolism with PBPK models of paracetamol and the potential glutathione-depletion biomarkers ophthalmic acid and 5-oxoproline in humans and rats". *Integrative Biology* 5 (2013), p. 877. 10.1039/c3ib20245c.

[39] M. P. Brigham, W. H. Stein, and S. Moore. "THE CONCENTRATIONS OF CYSTEINE AND CYSTINE IN HUMAN BLOOD PLASMA". *Journal of Clinical Investigation* 39 (1960), pp. 1633–1638. 10.1172/JCI104186.

[40] H. Kacser and J. A. Burns. "Rate control of biological processes". *Symp. Soc. Exp. Biol*. Vol. 27. 1973, pp. 65–104.

[41] H. V. Westerhoff and Y. D. Chen. "How do enzyme activities control metabolite concentrations? An additional theorem in the theory of metabolic control." *European journal of biochemistry* 142 (1984), pp. 425–30. 10.1111/j.1432-1033.1984.tb08304.x.

[42] S. Harvey Mudd *et al.* "Infantile hypermethioninemia and hyperhomocysteinemia due to high methionine intake: a diagnostic trap". *Molecular Genetics and Metabolism* 79 (2003), pp. 6–16. 10.1016/S1096-7192(03)00066-0.

[43] N. Swainston *et al.* "Recon 2.2: from reconstruction to model of human metabolism". *Metabolomics* 12 (2016), p. 109. 10.1007/s11306-016-1051-4.

[44] E. Noor *et al.* "An integrated open framework for thermodynamics of reactions that combines accuracy and coverage". *Bioinformatics* 28 (2012), pp. 2037–2044. 10.1093/bioinformatics/bts317.

[45] S. Sahoo, L. Franzson, J. J. Jonsson, and I. Thiele. "A compendium of inborn errors of metabolism mapped onto the human metabolic network". *Molecular BioSystems* 8 (2012), p. 2545. 10.1039/c2mb25075f.

CHAPTER 7

Simultaneous integration of gene expression and nutrient availability for studying metabolism of hepatocellular carcinoma

7.1 Introduction **220**
7.2 Results **222**
GENSI methodology 222
Metabolic genes: expression in two hepatoma cell lines 223
Converting RNA-seq data to RAS 227
Conversion of NA data into MUR 227
An FBA-based scaling methodology 228
Metabolic flux potential as predicted by flux variability analysis 229
Experimental verification 234
7.3 Discussion **237**
7.4 Materials and Methods **239**
Simulations 243
In vitro experiments 244
Supplementary Information **246**

Adapted from:
E. Węglarz-Tomczak, T.D.G.A. Mondeel, D.G.E. Piebes, H. V. Westerhoff, Simultaneous Integration of Gene Expression and Nutrient Availability for Studying the Metabolism of Hepatocellular Carcinoma Cell Lines, Biomolecules. 11 (2021) 490. 10.3390/biom11040490

> "A new scientific truth does not triumph by convincing its opponents and making them see the light, but rather because its opponents eventually die."
>
> — Max Planck [1][1]

Abstract

How cancer cells utilize nutrients to support their growth and proliferation in complex nutritional systems is still an open question. However, it is certainly determined by both genetics and an environmental-specific context. The interactions between them lead to profound metabolic specialization, such as consuming glucose and glutamine and producing lactate at prodigious rates. To investigate whether and how glucose and glutamine availability impact metabolic specialization, we integrated computational modeling on the genome-scale metabolic reconstruction with an experimental study on cell lines. We used the most comprehensive human metabolic network model to date, Recon3D, to build cell line-specific models. RNA-Seq data was used to specify the activity of genes in each cell line and the uptake rates were quantitatively constrained according to nutrient availability. To integrate both constraints we applied a novel method, named GENSI (Gene Expression and Nutrients Simultaneous Integration), that translates the relative importance of gene expression and nutrient availability data into the metabolic fluxes based on an observed experimental feature(s). We applied GENSI to study hepatocellular carcinoma addiction to glucose/glutamine. We were able to identify that proliferation, and lactate production is associated with the presence of glucose but does not necessarily increase with its concentration when the latter exceeds the physiological concentration. There was no such association with glutamine. We show that the integration of gene expression and nutrient availability data into genome-wide models improves the prediction of metabolic phenotypes.

7.1 Introduction

Cancer cells adapt their metabolism to promote fast cellular proliferation and long-term maintenance [2–5], thus facilitating the uptake and conversion of nutrients into biomass. Most metabolic signatures are shared across different kinds of cancer cells, including one of the best recognizable, namely changes in glucose metabolism that give rise to the Warburg effect [2–8], and an increase in biosynthetic activities (such as nucleotides, lipids, and amino-acid synthesis). However, the Warburg effect also plays an important role in many other cell types involved in immunity, angiogenesis, pluripotency, and infection by pathogens [9]. The Warburg effect is characterized by an increased rate of glucose uptake and part of glycolysis-derived pyruvate is diverted to lactate [2–10], which produces much less ATP per glucose than its oxidation to carbon dioxide would. This metabolic

[1] Reportedly framed by Otto Warburg and hung above his desk (`https://nyti.ms/24NFuTD`).

signature enables fast-dividing cells to satisfy anabolic needs for biomass production and is accompanied by a suppression of apoptotic signaling [11–14]. The high glucose consumption during the Warburg effect also provides a higher production of reduced nicotinamide adenine dinucleotide phosphate (NADPH2) through the pentose phosphate pathway, which provides electrons for cell proliferation [15]. However, it does mean that oxidative metabolism (i.e. respiration) is damaged. Instead, respiration and other mitochondrial activities are required for tumor growth [4, 16]. Many cancer cells also have an increased uptake of glutamine [17–21]. The partial catabolism of this glutamine to lactate by cancer cells has been called the WarburQ effect [22]. Some rapidly proliferating cells are particularly dependent on glutamine, and undergo necrosis upon glutamine depletion [20].

Genome-scale metabolic models (GEMs), which are at the core of some bottom-up systems biology approaches, can be used to predict cell physiology (e.g., growth rate and metabolic fluxes) under different conditions and improve our understanding of cell metabolism [23, 24]. Since GEMs utilize a Boolean formulation connecting genes to reactions, they have been used extensively as platforms for analyzing mRNA expression data to elucidate how changes in gene expression impact cellular phenotypes [25–37]. There are two fundamental approaches for integrating gene-expression data into GEMs: (1) based on direct integration of the gene expression information into the flux bounds and (2) based on a categorization of genes. The first way includes, for example, setting the fluxes to zero if an expression of their associated genes was low [27] and the maximum allowable flux value as a function of measured gene expression [28]. In the second approach the reactions are divided into different categories based on gene expression (e.g., highly or lowly expressed) and then reactions with highly expressed genes are associated with high flux and reactions with lowly expressed genes with non-high flux. Such an approach had been applied in the Gene Inactivity Moderated by Metabolism and Expression (GIMME) tool [29] by minimizing the flux through reactions whose associated genes' expression falls below a given threshold. In contrast, Shlomi et al., in their Integrative Metabolic Analysis Tool (IMAT), divided the reactions into those associated with highly expressed genes and those associated with lowly expressed genes and then maximized the number of reactions whose fluxes are consistent with their gene expression state [30]. Graudenzia et al. introduced the data integration framework named Metabolic Reaction Enrichment Analysis (MaREA) by projecting RNA-seq data onto metabolic networks by assigning a score for each reaction in the network (Reaction Activity Score, RAS). The score is calculated based on the expression of genes encoding for the associated enzyme(s) [31]. Such a methodology is highly useful due to the fact that metabolic reconstruction of higher organisms does not have a one-to-one relationship between genes and network edges, due to the existence of isozymes and protein complexes.

However, genetics is not the only determinant of metabolic phenotype. Cancer cell metabolism, similar to any other type of cell, is also influenced by metabolic constraints imposed by environmental and tissue-specific contexts [16, 38–48]. Numerous microenvironmental factors influence cancer cell

metabolism [38–40, 44, 48] including tumor acidity [49, 50] that is directly connected with lactate secretion and tumor nutrient levels [41, 42, 47]. In particular, environmental nutrient availability (NA) is an important regulator of cancer cell metabolism, and therefore an important environmental determinant of cancer cell metabolism is diet, which can affect the availability of nutrients within tumours [38, 39, 47, 51].

However, including concentrations of nutrients into GEMs in order to study their impact on cancer cell metabolism is not yet widely implemented. The existing methods for including nutrients concentrations into GEMs, which are based on the correlation of exchange fluxes with relative estimates of consumption and/or secretion rates, were applied to study the metabolic adaptation of Escherichia coli in complex nutritional systems [52, 53].

In the present study, we simultaneously constrain GEM and integrate linearly both: cell-intrinsic factors (gene expression level) and cell-extrinsic factors (nutrient availability). We applied a novel method, named Gene Expression and Nutrients Simultaneous Integration (GENSI), that translates genes expression and nutrient availability data into the fluxes through a scaling factor. The maximum allowable flux value of reactions associated with genes was constrained as a function of RAS [31] calculated based on gene expression level. While the maximum allowable flux value of the exchange reaction was constrained as a function of the possible consumption rate of the available nutrients. We integrated those constraints through Flux Balance Analysis [54] and a 'scaling factor'. Using GENSI we prepared specific models of the two hepatocellular carcinoma cell lines, HuH7 and PLC/PRF/5, and we used these to predict the influence of glucose and glutamine availability on proliferation rate and some cancer-related metabolic behaviors. We were able to identify glucose and glutamine (in)dependencies of both cell lines. Predictions were then confirmed experimentally.

7.2 Results

GENSI methodology

We designed GENSI as a method to integrate relative gene-expression and nutrient availability data into the human genome-wide metabolic reconstruction. Our purpose was reducing the solution space of optimal fluxes to provide results that can predict cell physiologies based on both cell-intrinsic and cell-extrinsic factors. Cancer cell metabolism, which aroused our interest, is also influenced by metabolic constraints imposed by environmental contexts. One of the important environmental determinants of cancer cell metabolism is diet, which can affect the availability of nutrients within tumours.

The GENSI workflow along with an illustration of the method performed on a GEM model is presented in Figure 7.1. GENSI requires three inputs: (1) a GEM model, (2) gene-expression data and (3) nutrient availability data (Figure 7.1).

The first step of the GENSI framework consists of pre-processing of the GEM model and includes: (1) preparing the model and includes blocking of the uptake

fluxes of various metabolites that are not present in NA, conversion of the gene identifiers that they are compatible with symbols used in RNA-seq data; (2) conversion of the gene-expression levels into RAS score [31]; and (3) conversion of the NA data into Maximal Uptake Rate (MUR) (Figure 7.1).

In the next step GENSI simultaneously translates RAS and MUR into the fluxes using a scaling factor of the RAS that is adjusted until experimentally observed features appear. As the direct correlation of gene expression data with maximal flux is not possible to apply in human metabolic map RECON3D [55], due to the existence of isozymes and protein complexes, we assigned to each reaction a RAS by summing over isoenzymes (for OR logic) and taking minima of subunits of a complex (for AND logic) of the TPM scores for the genes coupled to each reaction from the pre-processing step [28, 31, 56]. In this way, isoenzymes are thought to contribute additively to the activity of a reaction whereas the lack of even one subunit of an enzyme complex can bring down a reaction's activity [31]. To integrate nutrient availability data into the GEM model we propose using Maximal Uptake Rate (MUR). We defined MUR for all exchange reactions in the GEM. It describes the rate of maximum possible uptake over time f or substances (nutrients) available for the cells.

The GENSI framework was applied to integrate transcriptomic data from two hepatocellular carcinoma cell lines, Huh7 and PLC, and NA data. In our study, NAs were limited by the composition of the medium. We cultured both cell lines in six different conditions with various concentrations of glucose and glutamine. We call the NA data from the medium with 25 mM glucose and 4 mM glutamine concentration *NA1*, 25 mM glucose and 0 mM glutamine *NA2*,5 mM glucose and 4 mM glutamine *NA3*, 5 mM glucose and 0 mM glutamine *NA4*, 0 mM glucose and 4 mM glutamine *NA5*, and 0 mM glucose and 4 mM glutamine *NA6* (see Table 1 in Material and Methods). Based on the observed experimental differences in growth rate between the media and cell lines we found a scaling factor that translated RAS and MUR data into the fluxes that matched the observed growth rates.

We obtained different GEM-RAS-MUR models (see Figure 7.1) constrained by different combinations of MUR and RAS to investigate the effectiveness of NA and transcriptomic data integration in reducing the metabolic flexibility of the provided solutions. GEM-RAS-MUR is an integrated model obtained by incorporating NA data and RAS data into a GEM model, which is RECON 3D in this work.

Metabolic genes: expression in two hepatoma cell lines

For our two cell lines, the RNA-seq dataset published by Ma et al. [57] reports on 17,726 genes. We extracted the RNA-seq records for 2232 out of the 2248 metabolic genes that surface both in this data set and in Recon3D (see section Methods). Comparing the mRNA levels (in terms of TPM scores) for the subset of metabolic genes to the genome-wide mRNA levels, we observed that most metabolic genes exhibited a higher than average expression, with a median expression level of $\sim$15 compared to $\sim$1 genome-wide and a mean expression level

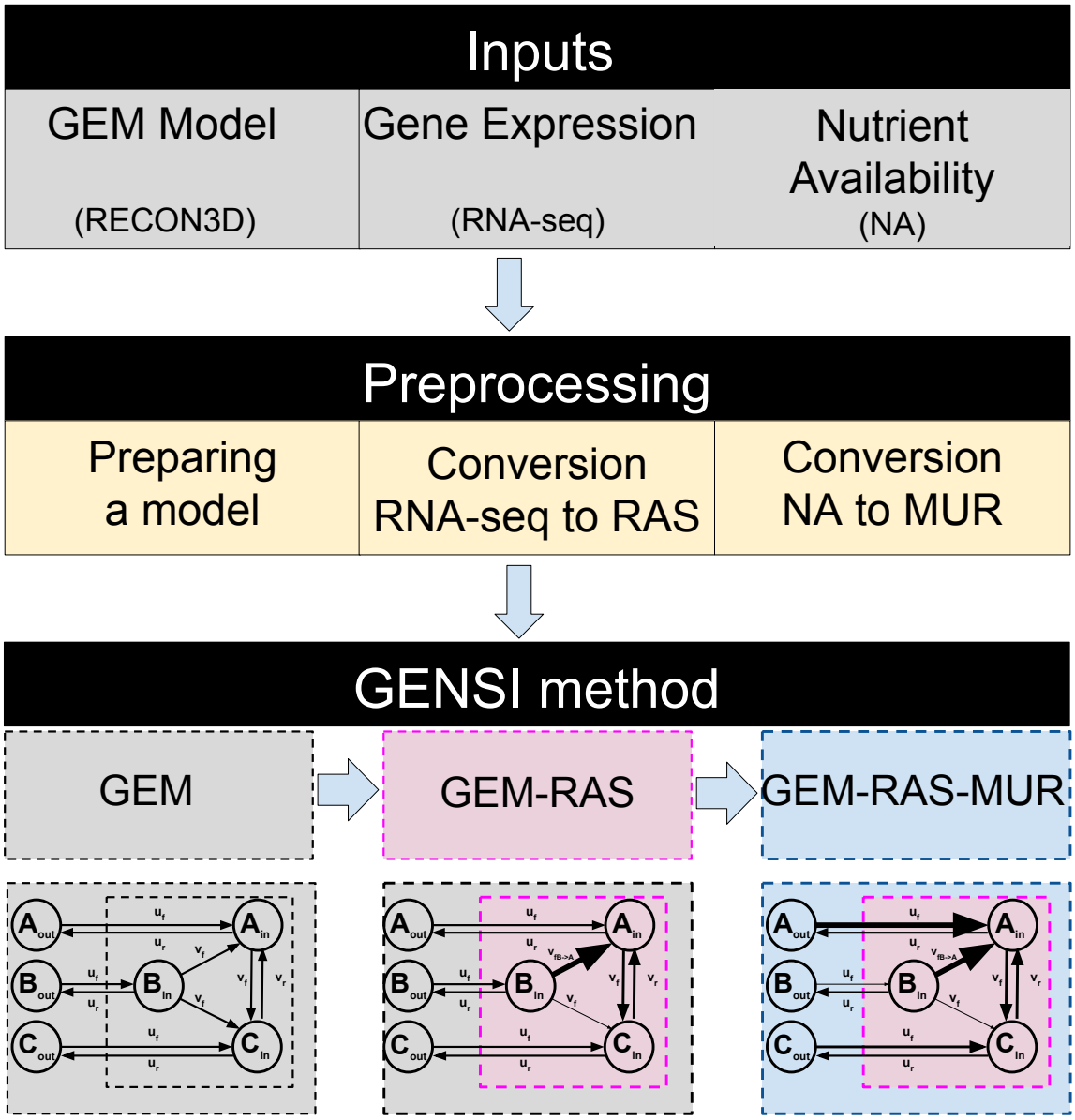


Figure 7.1: The Gene Expression and Nutrients Simultaneous Integration (GENSI) workflow requires three inputs: a Genome-Scale Model (GEM) model, gene-expression, and nutrient availability data (first block). In the pre-processing step (second block), the RNA-seq is converted into RAS [31] that is defined for any reaction r associated with gene(s) in the GEM and describes the extent of its activity, as a function of the expression of the genes encoding for the subunits and/or the isoforms of the associated enzyme(s). The Nutrient Availability (NA) data is converted to a Maximal Uptake Rate (MUR) that is defined for any exchange reaction r_{ex} in the GEM and describes the rate of the maximum possible uptake over time for substances available for the model (nutrients). Substances available for uptake are limited by the composition of the medium. Finally, GENSI (third block) integrates GEM, RAS, and MUR.

of ∼81 compared to ∼39 genome-wide.

We were confronted with a well-known issue with RNA-seq data integration in metabolic models, i.e. zero expression levels, including zeros in mRNAs encoding enzymes catalyzing essential reactions or combinations of enzymes that are essential. Of the 2232 metabolic genes, 262 and 310 had TPM scores equal to zero in Huh7 and PLC, respectively. Such zeros could be due to true absence or correspond to technical zeros where the gene had in fact been transcribed but was somehow not measured. Technical reasons may include inefficient cDNA synthesis due to tertiary structure formation, amplification bias, or low sequencing depth. Additionally, zeros may occur due to transcription bursting between somehow synchronized individual cells [58] or too small time-windows of expression. Given that an independently obtained set of microarray data might not suffer from quite the same problems, we assigned to genes with zero TPM scores in the RNA seq analysis, alternative TPM scores that reflected the microarray datasets (see Methods).

Huh7 and PLC being both hepatoma cell lines with a different history of oncogenesis, we further expected them to differ mostly in the expression of oncogenes and perhaps other genes involved in signal transduction or management of the genome, but not much in genes encoding metabolic enzymes. With respect to the metabolic genes, we expected both cell lines to be transcriptionally addicted to the same metabolic Warburg rewiring at the level of transcription. With this expectation, a plot of the expression levels of Huh7 versus the expression levels of the same genes in PLC should show all genes close to the diagonal line. Although for the majority of genes the correlation fell close to the diagonal, there were quite a few genes for which the expression levels differed between the two cell lines (Figure 7.2A).

We first looked at the genes that did correlate: In the RNA-seq dataset (prior to correcting for zeros as discussed above), 186 metabolic genes (∼8% of the 2232 genes) did not come with any transcript in either cell line. In total, 215 metabolic genes (∼10%) exhibited non-zero TPM values below the genome-wide median (∼1 TPM), again in both cell lines. Together, these constituted a common set of 401 metabolic genes that were expressed at a low level. On the high-abundance side, the two cell lines shared 1422 genes (∼63%) that were more highly expressed than the median gene. A 623-genes subset of these 1422 (∼28%) was commonly expressed above the genome-wide mean (39 TPM). These genes behaved in line with our expectation of metabolic similarity between these cell lines. They came however with a possibly important exception of some 425 genes that were off-diagonal in Figure 7.2A. In total, 415 metabolic genes (∼18%) exhibited a TPM score above the genome-wide median in both cell lines and a differential expression ratio of at least 3 (or below $1/3$). The data suggests that our expectation was not quite right: there was an appreciable metabolic difference between the two cell lines.

In accordance with the above, the expression ratio between the two cell lines was still mostly distributed narrowly around 1. The corresponding probability distribution was largely log-normal, but not quite: it had long tails on either side, suggesting that a disproportionate number of metabolic genes were much more

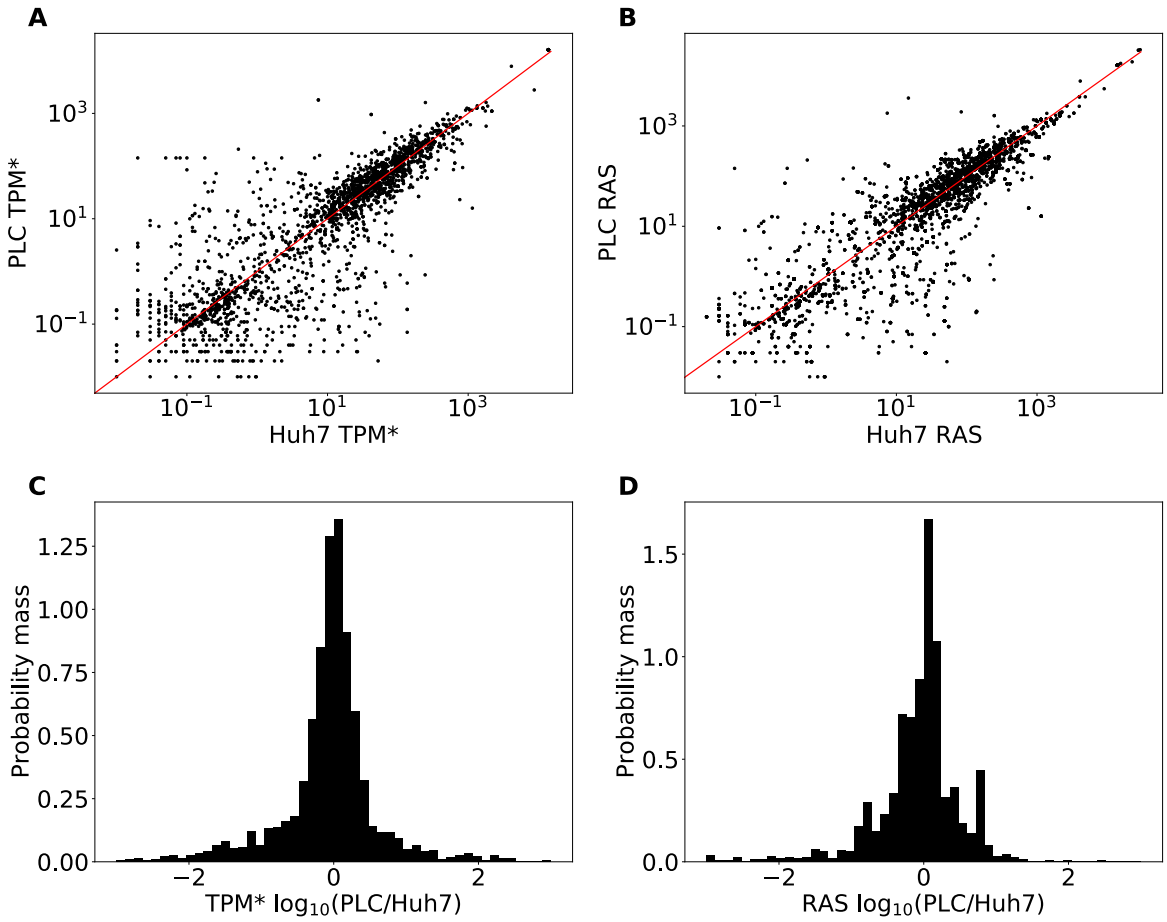


Figure 7.2: (**A**) Correlation of the mRNA levels (in terms of TPM* scores) between the Huh7 and PLC cell lines for all metabolic genes represented in Recon3D. The line PLC -TPM* = Huh7-TPM* represents theoretical identical scores between the cell lines. (**B**) 2D comparison of the RAS scores (dealing with multi-subunit proteins and isoenzymes) of the Huh7 and PLC cell lines for all metabolic genes represented in Recon3D. (**C**) Histogram of the probability mass function (PMF) of the log10 of the ratio of the TPM* scores between the two cell lines (PLC relative to Huh7) shown in (**A**). (**D**) Histogram of the PMF of the $\log_{10}$ of the ratio of the RAS scores between the two cell lines (PLC relative to Huh7) shown in (**B**).

expressed in Huh7 than in PLC and vice versa (Figure 7.2C).

Converting RNA-seq data to RAS reduces but does not eliminate metabolic differences between Huh7 and PLC

There are at least two reasons why the different expression of metabolic genes between two cell lines might not affect the activities of the corresponding biochemical reactions. First, the differences in expression level between the two cell lines could be in enzyme subunits that are abundant as compared to other subunits that are equally expressed. Second, the two cell lines may express different proteins (isoenzymes) that catalyze the same reaction. Recon3D dealt with this issue qualitatively through its gene-reaction coupling rules. We used a quantitative version of these rules to assign a Reaction Activity Score. The RAS for a metabolic reaction reflects the expression levels of isoenzymes and components of multicomponent complexes that may catalyze that reaction (Materials and Methods, and [28, 31, 56]). Out of the 10,601 reactions in Recon3D, 5938 reactions were assigned a RAS in this manner. For the 2999 reactions catalyzed by single genes, the RAS scores were taken equal to the TPM values. The remaining 4663 reactions are not linked to any genes: they represent so-called 'exchange reactions' between the cells' immediate environment and the outside world, or non-enzyme-catalyzed reactions and transport within the cells or across their membranes. We left the bounds of such geneless reactions unlimited at ± 1000.

We then asked whether the metabolic differences we found between the two cell lines would disappear when correcting for these isoenzyme and enzyme subunit issues by assigning reaction activities. In Figure 7.2B, we correlate the RASs between the two cell lines, and panel 1D shows the distribution over the metabolic genes, of the RAS ratios between the two cell lines. The RAS correlation between the two cell lines is only a little stricter than that of the individual mRNAs. The standard deviation in the RAS ratios is 14 compared to a standard deviation of $\sim$177 for the TPM* ratios. The fraction of outliers outside the lognormal distribution of the ratio remains substantial, however. Supplementary Excel Table S3 lists the 92 outlier reactions with a log10 RAS ratio > 2 or < -2.

Conversion of NA data into MUR

In metabolic networks like RECON3D, all metabolites with defined exchange reactions can be taken up and secreted, i.e. compounds enter and exit the extracellular environment via 'exchange' reactions. The GEM is not able to import compounds unless an exchange reaction from the external environment to the inside of the cell is present. To predict the effects of nutritional differences in terms of all components in the medium including the various concentrations of glucose and glutamine, between our six different medium conditions (see Table 1 in Material and Methods), on the global metabolic behavior of the two cell lines we first converted nutrient availability data into maximum uptake rates (MURs) that describe the rate of maximum possible uptake over time for substances present in the medium. We defined the MUR for the exchange reaction r_j as the absolute

value of the difference in the concentration of the substate s_j in the extracellular environment (see methods) over time. In this work, we assume that all amounts of each substance can be uptaken, and therefore the concentration of the substate s_j in time 48 h is zero. We equated the upper bound of each uptake reaction to the corresponding MUR, setting the bound to zero if the metabolite was absent from the medium. Export was left unlimited for all metabolites that were allowed to be exported in the default Recon3D map (see Supplementary Excel File S2 for the list of such metabolites). As a consequence of this approach the unit of the uptake fluxes is equal to the unit used for the MUR, i.e. concentration deviated by time. In our study, we set the time to two days (48 h) due to experimental conditions.

An FBA-based scaling methodology

In the GENSI framework, we proposed a variant of the approach published by Graudenzia et al. [31]. We set flux bounds of reactions associated with genes proportional to RAS scores, i.e. $b_i = \alpha \cdot RAS_i$, where b_i is the flux bound on reaction i, is a factor independent of the reaction identity, and RAS_i is the RAS based on the expression levels of an enzyme(s) catalyzing reaction i. When reaction i is reversible its forward flux is bounded by b_i and its backward flux is bounded by $-b_i$. When reaction i is irreversible the flux bound in the impossible direction remains zero and the flux bound in the possible direction is set to b_i. Essentially this approach assumes that the V_{max} of any enzyme is proportional to the corresponding mRNA transcript level. Our variant of the MAREA approach allows us to scale the RAS data, in reference to a specific dataset of NA (see above).

It is a priori unclear how large the factor should be. There are two possible sources of limitation for the model: the MUR (as represented by maximal uptake rates) and the RAS (consequent to transcriptomics). We here wish to examine the case where the enzyme expression levels begin to impose limitations on the model output. We perform FBA analysis [54] starting with high α values, where the enzyme expression levels are not limiting, and maximal biomass flux is determined by substrate concentrations in the medium. Then we decrease α and hence all resulting metabolic flux bounds uniformly until we see differential effects on the predicted maximal growth rates across media conditions. Where this occurs significantly, we fix α, which thereby becomes a fitted parameter.

We analyzed RAS with the focus on metabolic differences between the cell lines by computing the steady-state flux pattern for the maximal biomass synthesis flux for each medium and for each cell line across a range of values for the factor $\alpha \in [3 \times 10^{-4}, 1.6]$ (Figure 7.3). For either cell line Figure 7.3A shows that for factor α in excess of 0.5, the predicted growth rates differed between media conditions, the ones at 25 mM glucose being about double those at 5.6 mM glucose. Furthermore, dependence on glutamine concentrations was predicted, but this dependence was smaller.

When decreasing the factor alpha to below 0.5 the first effects of the transcriptomics start to be noticeable, i.e. growth predictions for the two cell lines start to diverge in some medium conditions. When decreasing the factor α to below 0.004, the MUR dependence disappeared for four out of six modeled medium

conditions: the same biomass flux was then predicted which was still significantly higher than that for the remaining two medium conditions at zero glucose (5 and 6). Huh7 predictions for MUR 1–4 converged already at higher values for α than did the predictions for PLC. Huh7 was predicted to have equal or lower growth rates than PLC across all conditions. MUR 1, followed by MUR 2, was predicted to yield the highest biomass fluxes for the high values of α, corresponding (see above) to the absence of gene-expression limitations. This reflects the model's sensitivity to carbon input for high α values since medium 1 and 2 contained the highest levels of glucose. $\alpha = 0.004$, the simulations for MUR 1–4 yielded equal biomass fluxes which were still larger than the predicted fluxes for MUR 5 and 6 which lack glucose. This indicates that by reducing the factor α, the model can be made more (high α, hence no limitation by low transcription of metabolic genes and thereby limitation by uptake) or less (low α, hence strong limitation by low transcription of metabolic genes) sensitive to variation in concentration of the growth substrate.

The fact that MUR 1-4 converges to similar biomass synthesis flux optima for low levels may reflect a shared limiting reaction downstream of (and at a flux bound smaller to the bound of) the exchange reaction the flux bound of which keeps monotonically decreasing with decreasing α. In Figure 7.3B we summarize the predicted maximal biomass fluxes for the factor $\alpha = 0.1$, which is at the transition between limitation by extracellular substrate levels and intracellular expression levels. It shows that reduction of glucose concentration does not decrease maximal biomass flux as we observed in experiments (compare Figure Figures 7.3B and 7.3C).

We observe that nontrivial predictions for limitations imposed by medium composition and gene expression can be computed by GENSI method, such as that (i) both in the absence and in the presence of glutamine the growth rate should be independent of glucose concentrations between 5.6 and 25 mM, yet decrease appreciably in the absence of glucose, (ii) the specific growth rate of Huh7 cells is lower than that of PLC cells, (iii) in the absence of glucose, the cells should be able to grow on glutamine, but (iv) growth rate on glutamine alone should be much lower than on glucose alone.

Metabolic flux potential as predicted by flux variability analysis

FBA is oblivious of metabolic regulation other than that it philosophizes about what flux should be optimal for the cell in view of some objective. The transcriptome and extracellular-concentrations informed flux bounds that we here implemented, merely define ranges of the fluxes rather than that they precisely predict the fluxes. Moreover, fluxes through intracellular biochemical reactions are also determined by metabolic regulation [59], i.e. by the concentrations of intracellular metabolites. For the precise predictions of fluxes, one needs fully dynamic models [60]. However, the kinetic information required for this approach is missing for mammalian cells.

We hereby can only predict the ranges of fluxes that are consistent with transcriptome and extracellular nutrient concentrations and this is done here by flux

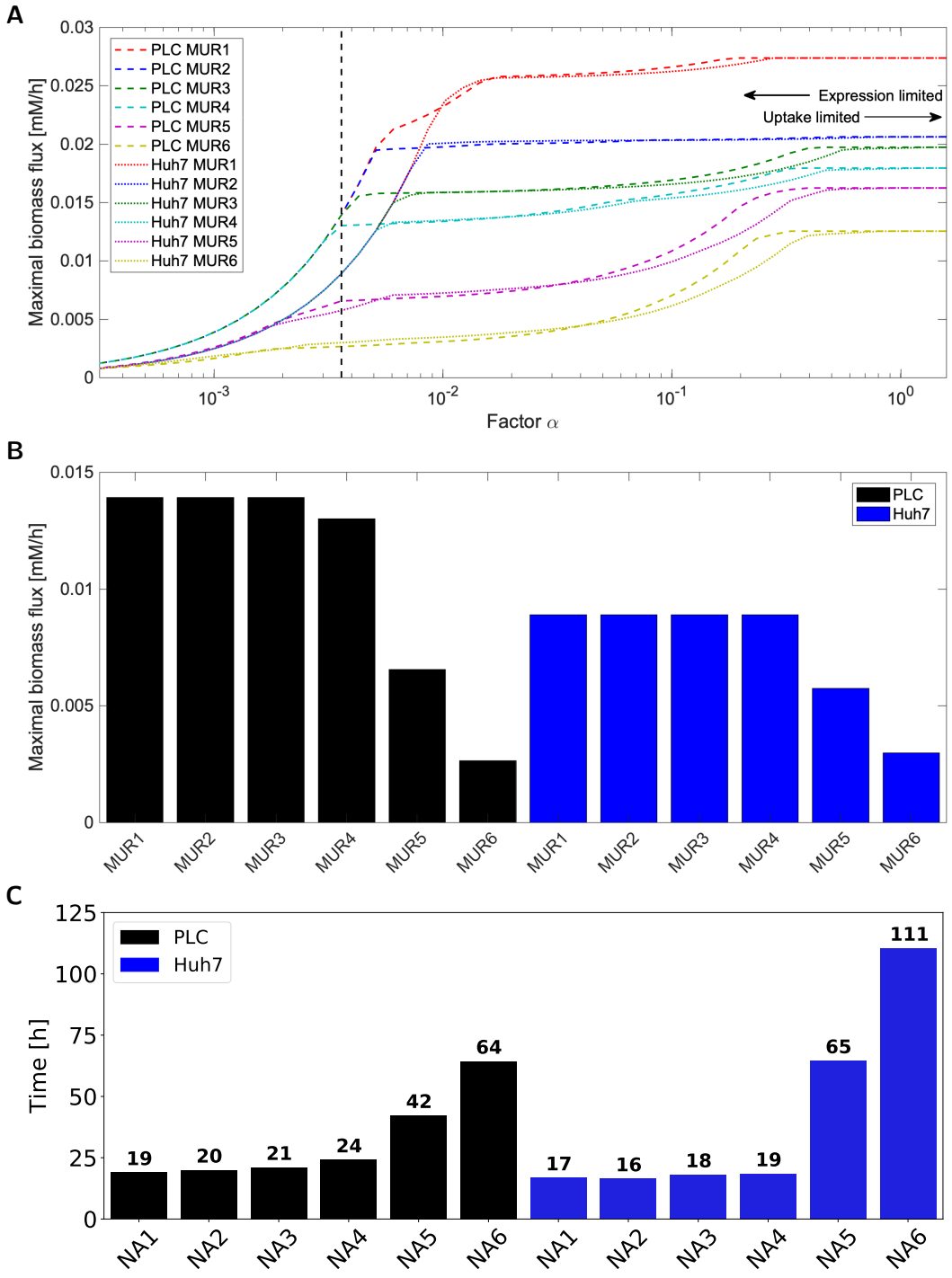


Figure 7.3: (**A**) Maximal in silico biomass flux predictions for Huh7 (dashed) and PLC (dotted) on six MURs versus the factor used to convert RAS to flux bounds. The labels MUR1-6 correspond to medium conditions listed in Table 7.1 and Table S7.4 and were effected as proportional uptake bounds. The dashed black line indicates $\alpha = 0.0036$ at which the maximal biomass flux predictions for MUR1-4 become virtually equal for both cell lines. (**B**) Maximal biomass flux predictions for PLC and Huh7 across the six GEM-RAS-MUR models. (**C**) Cell doubling time calculated from a linear equation describing the change in cell number over time for PLC and Huh7 across the six different media conditions.

variability analysis (FVA) [61] to test the prediction of the GEM-RAS-MUR models. We performed FVA analysis for allowed exchange reactions while supporting the biomass production rate. For the factor value $\alpha = 0.0036$, indicated by the black vertical line in Figure 7.3, we analyzed each of the 12 GEM-RAS-MUR models in terms of the possible ranges its production fluxes of lactate, pyruvate, ATP and CO_2 and its uptake fluxes of glucose, glutamine, oxygen, and phenylalanine. Here we maintained the maximal biomass flux for each specific medium and transcriptome (i.e. the biomass fluxes listed in Figure 7.3, which differ between the 12 GEM-RAS-MUR models) (Figure 7.4). The results for lactate and pyruvate production are non-negative since these compounds are not in the growth medium and thus cannot be taken up. In order to avoid thermodynamically infeasible ATP synthesis, the ATP hydrolysis reaction was non-negative by design. The fluxes in Figure 7.4 for glutamine and CO_2 can be both negative and positive due to these compounds being present in the medium. If the lower end of its bar in Figure 7.4 is positive, that compound must be produced for the cell to grow at maximal growth rate, i.e. it is a primary metabolite. When the upper end of the bar is negative it indicates that the compound needs to be taken up for maximal growth.

In Figure 7.4, we see that in the media where glucose is present (NA1-NA4) some glucose uptake is essential for attaining the maximal growth rate. In NA5-NA6 glucose uptake is always zero due to its absence from the medium. Phenylalanine, an essential amino acid, functions as a positive control. Its uptake proves indeed essential in all media for both cell lines as expected. In most media, its uptake rate can vary from the amount of phenylalanine in biomass to 0.4, the rate at which it can be used to provide nitrogen to other parts of anabolism. This maximum uptake rate of 0.4 [mM/2d] is the same for all media and both cell lines, reflecting that this corresponds to its concentration in the media. In the absence of glutamine and in the presence of low glucose (NA4), PLC needs to make full use of this phenylalanine in order to achieve its maximum growth rate, but Huh7 cells could still vary the amount of phenylalanine used whilst attaining the same growth rate. Maximum lactate, pyruvate, and ATP production capabilities track the total amount of carbon in the medium within each cell line, with subtle differences between the two cell lines. The gene expression levels appear to be consistent with shifting to virtually complete metabolism of glucose and glutamine to lactate whilst maintaining the maximum growth rate. At the same biomass production flux, lactate efflux could also be as high as 55 mM/2d, roughly corresponding to 2 lactate per maximum glucose consumed plus one lactate per maximal glutamine consumed. However, in all models, the lactate secretion can also be zero while maintaining maximal growth. This shows that glucose conversion to lactate can vary greatly, and may also reflect that in our models the cells can produce and secrete other compounds such as pyruvate. The in silico cells are not addicted to the Warburg effect.

Glutamine uptake is only essential for both PLC and Huh7 in medium 5 and for PLC only in medium 3. In medium 3 for PLC it is then required at a very low amount to achieve optimal growth (as indicated by the upper end of the bar) whereas in medium 5 in both cell lines the maximal uptake bound has to be hit to achieve maximal growth (see the markers at -4). Because (in silico) the maximum

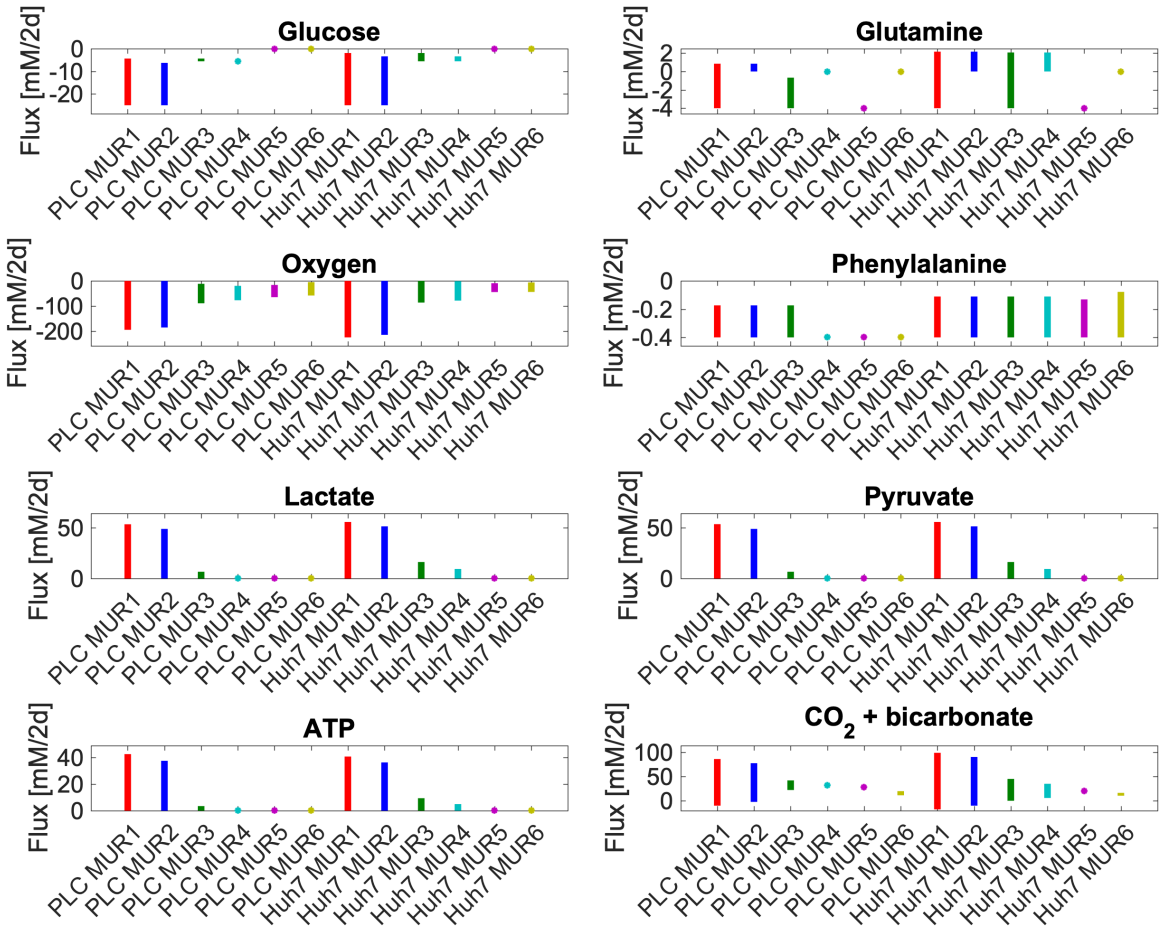


Figure 7.4: The range of uptake (if negative) and secretion (if positive) fluxes of various metabolites: computed to be consistent with maximal biomass flux and steady-state. Since we set maximal uptake fluxes equal to medium concentrations these rates differ from reality by some undetermined factor which is identical for all conditions. Minimal and maximal exchange fluxes for each of the compounds lactate, phenylalanine, glutamine, pyruvate, oxygen (O_2) and CO_2 + bicarbonate and ATP hydrolysis flux (i.e. ATP + H_2O => ADP + P_i + H^+), were calculated using flux variability analysis [62] while requiring the model to produce the same maximum possible biomass fluxes shown in Figure 7.3. Maximal glucose and glutamine uptake fluxes had been set to their medium concentrations divided by 2 days (see Table 1 in Material and Methods). In these calculations, ATP is treated differently from the others. For the others, the reaction was and remained present and potentially carrying varying flux when any of the yet other fluxes was manipulated in the FVA. For the ATP, the ATPase reaction was absent (no growth rate-independent maintenance, therefore) when doing FVA for any of the others and only present when ATP synthesis was manipulated by forcing flux through an ATP hydrolyzing added reaction. For visibility, we used starred markers to indicate small flux ranges. These markers are typically located at minimal or maximal flux boundaries, e.g., zero glucose uptake in NA5 and NA6 or maximal phenylalanine uptake in PLC NA4. Uptake and secretion fluxes of metabolites have units of mM/2d.

growth rate of Huh7 is lower it has the luxury of producing glutamine from glucose whilst growing maximally in M3 whereas this is not possible for PLC. Both cell lines may produce glutamine in M1 and M2 owing to the excess glucose in those media. We conclude that the model cells are insensitive to glutamine concentrations in the medium in the presence of high glucose but glutamine-sensitive in the absence of glucose.

In all models, a small amount of oxygen must be taken up. CO_2 (either as CO_2 or as bicarbonate) may either be produced or taken up in M1-M2 for both cell lines and M3 for Huh7 and may only be produced in M3 for PLC and M4-M6 for both cell lines. CO_2 uptake might have to do with the reversal of the isocitrate dehydrogenase reaction, which produces isocitrate as a substrate for ATP citrate lyase producing cytosolic acetyl CoA for lipid and cholesterol synthesis [27]. For conditions where CO_2 or bicarbonate production is required, this may point to oxidative phosphorylation being required for maximal biomass production. Oxygen uptake was essential even in conditions where oxidative phosphorylation (interpreted as CO_2 production) seems not to be required. In these cases, oxygen uptake may be necessary for the synthesis of tyrosine, cholesterol, and other lipids that are part of the biomass definition and absent from our growth media. We checked that removing cholesterol from the biomass equation reduced the need for oxygen, but it did not remove it.

The possibility to grow at a maximum rate in the absence of CO_2 production in some conditions highlights the possibility for cells to obtain all the Gibbs energy they need for maximal growth only from the conversion of glucose to lactate. This may underlie the selection of the Warburg effect by a-social cells. It does not quite correspond to the Warburg and WarburQ effects, however: the in silico cells are not addicted to the absence of respiration, as they can still respire all this substrate whilst growing at the same rate. In PLC cells, but not in Huh7 cells, the maximal growth rate in low glucose medium without glutamine (MUR4) does require oxidative phosphorylation, consistent with the glutamine to lactate pathway elucidated by Damiani et al. [22]. These and other apparently minor differences between cell lines in our FVA results are of interest, as they suggest that drugs, in this case, ones that inhibit respiration, should be effective against some cancer cells and not others, also depending on extracellular metabolic conditions.

We further explored this by plotting some essential reactions for respiration to occur analogously to Figure 7.4 in terms of their minimal and maximal possible flux allowed while maintaining maximal biomass flux for the medium and cell type specified (Figure 7.5). In media, with glucose, the maximum growth rate does not require flux through cytochrome oxidase and oxidative phosphorylation with the exception of NA4 for PLC. Figure 7.5 suggests that for PLC respiration in terms of flux through cytochrome oxidase is required for maintaining the maximal biomass flux in media 4–6 whereas for Huh7 this is required for media 5 and 6: rather than a range of fluxes, a precise non-zero flux magnitude is required. This suggests that only in those cases of limiting metabolic substrate, the maximal growth rate depends strictly on ATP produced by oxidative phosphorylation. This is in full agreement with the interpretation of the CO_2 + bicarbonate panel in Figure 7.4. In all other cases, respiration is optional for maximum biomass

synthesis flux, suggesting that the cells can obtain their ATP from other processes including aerobic glycolysis. To maintain their maximum growth rate at the 5.6 mM glucose concentration, they do need to use virtually all that glucose, however (Figure 7.4).

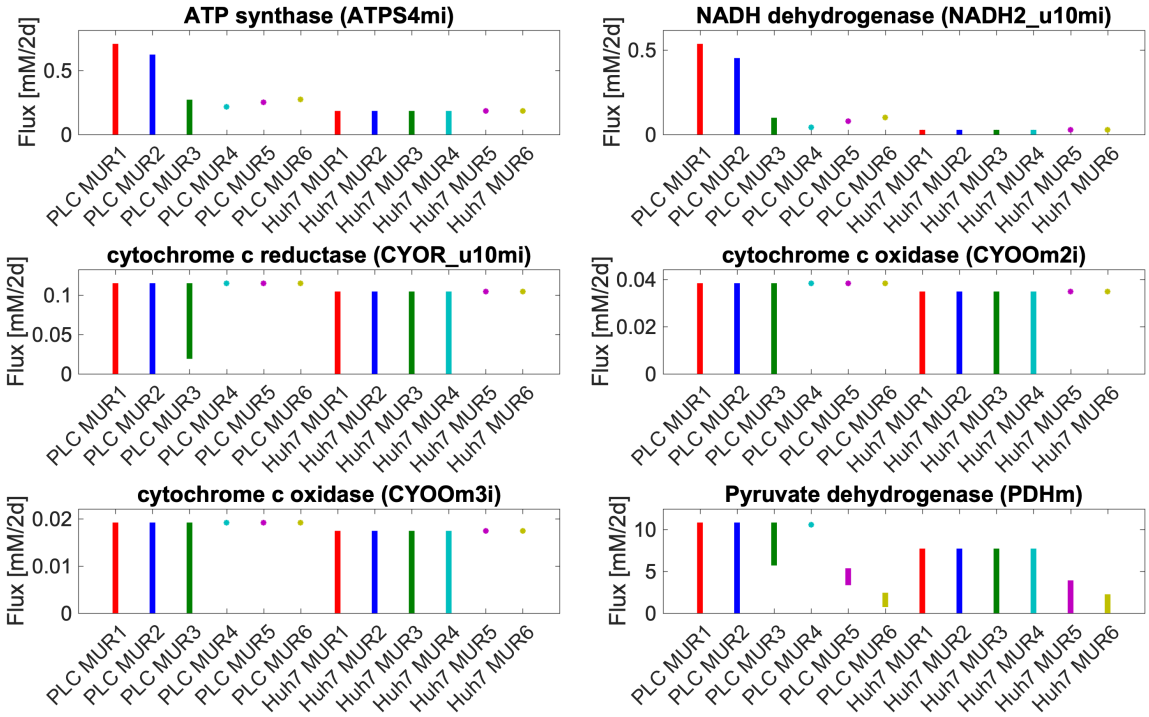


Figure 7.5: Range of allowed flux values through various reactions related to mitochondrial oxidative phosphorylation while maintaining the maximal biomass flux for the medium condition and cell type specified on the abscissa. See Table S7.3 for the detailed reactions. Uptake and secretion fluxes of metabolites have units of mM/2d.

Because of its assumptions of maintenance of maximum growth rate and the full capability of the network to allow for various fluxes, flux variability analysis makes few predictions that may be put to the test in this study. Exceptions are (i) both cell lines should be capable of consuming the 5.6 or 25 mM glucose offered to them, (ii) they are not addicted to a 100% aerobic glycolysis, but can reduce lactate production without giving up their maximum growth rate, (iii) at glucose concentrations around 5 mM they would make use of all that glucose to grow maximally, (iv) They should be capable of catabolizing glutamine both in the absence and presence of glucose.

Experimental verification

To verify whether GEM-RAS-MUR models obtained via GENSI give reliable predictions, we measured the concentration of the crucial components in the medium during culturing. We observed a decrease in the concentration of glucose with time such that some 5 mM was consumed by the PLC cells (Figure

7.6B). In the PLC cultures that started with 5.6 mM glucose, this resulted in almost full glucose depletion after the first day and night. In the case of the Huh7 cell line, the consumption of glucose was higher in rich glucose medium than in low glucose medium reaching some 8.5 mM and 5 mM, respectively. The further velocity of glucose consumption for Huh7 was the same, while in the case of PLC cells, after 24 h of culturing, stabilization appeared and lasted for the next 20 h (Figure Figures 7.6B and 7.6C). We did not observe a difference in glucose consumption between media with and without glutamine. With respect to glutamine consumption by the cells, we found that the level of glutamine did not change maintaining the 4.5 mM level for media with glutamine (that contained originally 4 mM) and 0.5 mM without.

We additionally measured the production of lactate and pyruvate by cells using the same media conditions as above. In the case of the PLC cell line, the production of lactate was the same during the first 24 h for NA1-NA4, the lactate concentration reaching some 5 mM at the end of this time period (Figure Figures 7.6C and 7.6D). After 55 h the concentration of lactate was twice higher in samples derived from the glucose-rich NA1 and NA2 (8.5 mM) than in samples taken from the low-glucose (5.6 mM) media NA3 and NA4, ostensibly because after 24 h the glucose had run out. In media low in glucose (NA3 and NA4), slightly over 40% of the glucose was consumed during the experiment was transformed to lactate, while for M1 and M2 this was a bit less, i.e. 35%. For the Huh7 cells, the trends were similar in the case of NA1–NA3. In the case of NA4, such production was achieved only after 45 h. In NA1 and NA2 lactate increased linearly with time attaining 10 mM at the end of the experiment.

Pyruvate production was quite small for PLC cells, from 0.15 mM (NA1 and NA6) to 0.4 mM (NA4) during the first 24 h. Pyruvate production was higher for media without glutamine (NA2, NA4, and NA6 vs. NA1, NA3, and NA5). For PLC cells, the lactate/pyruvate ratio (L/P) showed a positive relationship with the glucose consumption, being highest for NA1, and lowest for NA5. For Huh7 cells, the time-series for pyruvate (Figure 7.6F) look similar to those that show lactate secretion (Figure 7.6D). On NA1–4 the ratio of lactate to pyruvate changed over time from 6 to 10 whereas it stayed roughly constant for NA5 and NA6.

Consistent with the hypothesis that both cell lines exhibited a Warburg effect, they produced an amount of lactate proportionate to the amount of glucose that was present initially. Apparent differences merely derived from the glucose running out after some 30 h in the experiment starting at only 5.6 mM of glucose. Only part of the glucose was transformed to lactate, around $\sim$43% for low glucose media and perhaps a little as $\sim$33% in glucose-rich media. Thus, lactate production corresponded to only part of what might have been expected for full glucose conversion to lactate. This deficit in lactate secretion can be explained by the utilization of glucose for oxidative phosphorylation and as a source of carbon for the new biomass. Consistently, we did not observe any difference between NA1 and NA2 for both cell lines nor did we observe such differences between NA3 and NA4 for PLC; for Huh7, small deviations were noticed. For the two cell lines we examined, this proves the independence of lactate production from glutamine access.

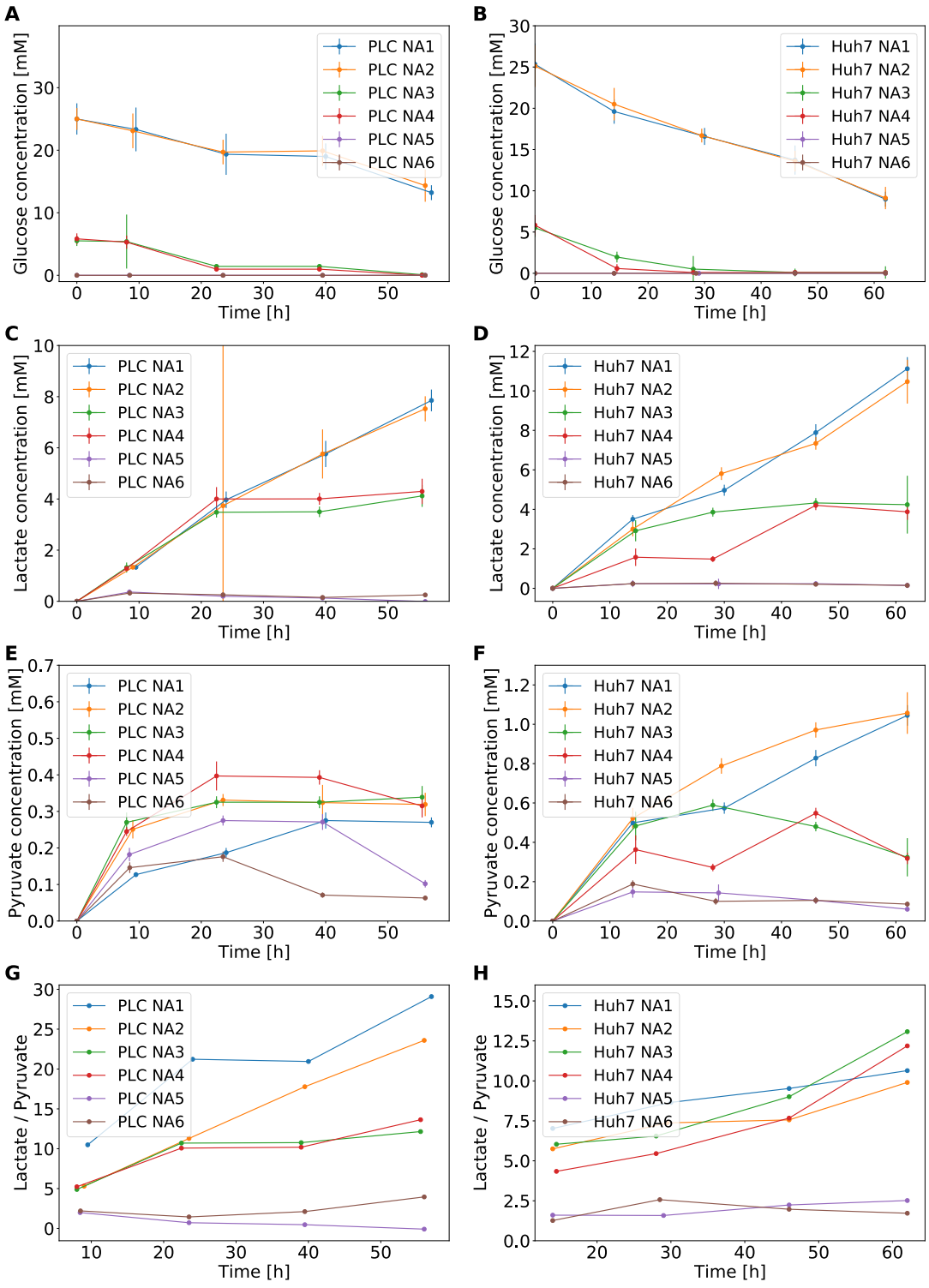


Figure 7.6: Metabolic performance as a function of time for Huh7 and PLC cells in six different media. Glucose consumption by PLC (**A**) and Huh7 (**B**) in four different media. In media M5 and M6 glucose was not detected. Lactate production by PLC (**C**) and Huh7 (**D**) cells. Pyruvate secretion by PLC (**E**) and Huh7 (**F**) cells. Ratio of extracellular lactate to pyruvate in PLC (**G**) and Huh7 (**H**) cells starting from the second time point measured in panels (**D**–**G**). In all six media the concentration of lactate and pyruvate before the experiments was 0 mM.

7.3 Discussion

In this work, we follow recently exploring interest on the impact of the diet on cancer cell metabolism and progression. We explore the idea that the utilization of nutrients by cancer cells is determined not only by cancer genetics but also by the metabolic environment.

To investigate whether cancer progression may be mediated through changes in the access to nutrients, genetics and the availability of nutrients have to be employed simultaneously. Genome-scale models that integrate all known metabolic reactions occurring in an organism into a single map give us such an opportunity. The gene-reaction coupling rule and exchange reactions allow for the integration of gene expression and nutrient availability data, respectively. Here, we aim to provide an FBA-based framework, named GENSI, for studying the influence of the nutrient availability on the rate of proliferation and metabolic phenotype based on transcriptomic and NA data. Our method could be applied to the study of cell metabolism of every cancer type and therefore could lead to a better understanding of how diet impacts cancer cell metabolism and identifying how different cancer types respond to different nutrients composition. GENSI translates the relative importance of gene expression and nutrient availability into the fluxes and generates high-quality GEM-based specific models (GEM-RAS-MUR) that can be used to predict metabolic changes upon nutrition shift by flux balance analysis.

Further, we applied the proposed method to study the influence of the glucose and glutamine availability on their consumption, some metabolic signatures, and the growth rate of the hepatocellular carcinoma cell lines PLC and six for Huh7. We used transcriptomic data from the cell population in its entirely and the models are a representation thereof. Single-cell transcriptomic studies suggest that cancer cell populations are heterogeneous, but single-cell metabolomics does not yet enable us to examine the consequences of metabolism. Thus, we assumed in the model that the cell population was homogeneous.

Our findings offer support for the predictive potential of genome-scale metabolic maps together with transcriptomic data sets and nutrient availability data. The predictions address the carbon and energy metabolism of cancer cells. Because much of this metabolism is essential for cell survival, the potential may translate to new drug targets in a long-neglected area of drug discovery. Indeed, we have shown that dependencies on nutrient availability and gene expression can be computed and that the results are then relevant enough to be compared with experimental work.

Notwithstanding these successes, our methodology comes with a number of issues. One of these relates to the translation of the expression level information to flux. In contrast to several methods developed to extract context-specific models [32, 33, 35] that focused on threshold selection with exception of some essential metabolic functions that are needed for cell growth, our method used a linear relationship between flux bounds and the transcriptome. Such a linear approach was first proposed by Colijn et al. [28] and applied in the MaREA framework by Graudenzi et al. [31], where it was shown to be of use in comparing

the metabolism of samples in distinct subgroups. The approach assumes that the V_{max} of a reaction is proportional to the level of the mRNA encoding the enzyme, with a proportionality constant equal for all reactions. It thereby neglects differential translation and posttranslational regulation, and assumes all k_{cat} to be equal. Furthermore, we assume that the cell lines in our experiments have not significantly evolved compared to those used in the transcriptomics datasets [57]. We did not take into consideration epigenetic and gene-expression changes that could occur during culturing. In addition, our approach equated MUR data (that we obtained based on NA data) with exchange reaction lower bound, again treating all compounds equally. The limitations of this step include the failure to take into consideration the kinetics and expression levels of transporters. What is also peculiar in our approach is the arbitrary magnitude of the ratio of the proportionality constant relating RAS level to enzyme bound to the proportionality constant relating MUR to exchange bound. We mediated this problem by setting the factor α to a value that enabled the models to shift from limitation by nutrient availability to limitation by expression level.

In silico experiments on GEM-RAS-MUR models showed that the specific growth rate of Huh7 cells is lower than that of PLC cells, which was also observed during experiments on these cell lines. Given the positive correlation between the in silico biomass flux and the experimentally determined growth rate, we speculated that the in silico results pertaining to the range of uptake and secretion fluxes of various metabolites might also correlate with experimental results. Modeling showed that in the absence and in the presence of glutamine the growth rate should be independent of glucose concentrations between 5.6 and 25 mM. Furthermore, experimentally the growth rate reached the same level for four nutrient availability (NA1–NA4) that contained 5.6 and 25 mM of glucose for both HCC cell lines. Moreover, both approaches, computational and experimental, showed that in the absence of glucose, the cells should be able to grow on glutamine but with a much lower growth rate than on glucose alone.

According to the flux variability analysis predictions, both cell lines should have a higher potential; they should be capable of consuming the 5.6 or 25 mM glucose offered to them. Yet, our observations indicated that PLC cells confronted with the 25 mM did not consume more than the around 5 mM during the first 24 h. They did not make use of that full potential. We conclude that there is a maximum amount of glucose per unit time that the cells 'wish to' handle at glucose levels above a few mM, because other issues than energy and carbon may limit the cells' growth. We use the term 'wish to' to indicate that this may be an issue of metabolic regulation: the gene expression would allow for higher glucose uptake fluxes.

The cells that had consumed in the first 24 h some 5 mM of the 5.6 mM glucose they had been incubated with, continued to grow for the next 24 h at virtually the same growth rate; they must have done this at a much-reduced glucose consumption rate. Since they also stopped producing lactate, we suspect that the small amount (approximately 0.5 mM) of glucose left provided the cells with sufficient ATP to drive their continued growth. The cells may have reverted from lactate production to respiration with its more than 15-fold higher ATP yield. This

highlights that there may be a limitation to these cells' addiction to lactate production: these cancer cells can shift to glucose oxidation. We observed a slightly different effect in the case of the Huh7 cell line: the consumption of glucose was almost two times higher in rich glucose medium (25 mM) than low on (5 mM) during the first 24 h. In silico analyses predicted and experiments confirmed that both cell lines, at glucose concentrations around 5 mM, made use of all available glucose to grow maximally. They also showed that upon glucose depletion and if asked to grow at the maximal growth rate, the cells should shift to glucose respiration (Figures 7.4 and 7.5).

Significant progress has been made toward developing the tool to study how diet and nutrition affect cancer progression. We obtained GEM-RAS-MUR models for the hepatocarcinoma cell lines and used them to predict cell growth and metabolite consumption/production. Our methodology could be applied to different types of cancer to investigate how cells respond to the diet and how to mediate these responses. Potential future applications of the GEM-RAS-MUR model would be to predict cell addictions and metabolic requirements in order to modulate them for dietary aid in cancer treatment. Moreover, GENSI methodology is not limited to cancer research, it could be applied to study the metabolism of any healthy and/or sick cells to study how diet and nutrition affect metabolic phenotype, proliferation, and functions of the cell.

7.4 Materials and Methods

GEM Model. The most comprehensive genome-scale model of human metabolism RECON3D [55] that includes information on 3288 open reading frames that encode metabolic enzymes catalyzing 13,543 reactions on 4140 unique metabolites, was used in this study.

RNA-seq data. RNA sequencing data for both cell lines was obtained through NCBI's GEO database. Specifically, the RNA-seq dataset from [57] was accessed from: `https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE86602`, accessed on 7 November 2018. The RNA-seq dataset pertains to cells grown in DMEM containing high glucose (Gibco BRL, Grand Island, NY, USA), 10% heat-inactivated fetal bovine serum (Gibco BRL), 100 mg/mL penicillin G, and 50 Âµg/mL streptomycin (Gibco BRL) at 37 °C in a humidified atmosphere containing 5% CO_2 [57]. The RNA-seq data records transcript levels in terms of a TP(K)M value (Transcripts Per Kilobase Million), i.e. read count normalized by gene length in kilobases (RPK) and divided by 1 million. Microarray transcriptomics data. We obtained microarray transcriptomics data from the MERAV database [63] (`http://merav.wi.mit.edu/`, accessed on 7 November 2018) for both the Huh7 and PLC/PRF/5 cell lines. There was data available from two experiments for Huh7 and one for PLC/PRF/5. All MERAV microarray datasets were renormalized together [63]. For the two Huh7 experimental datasets we averaged the signal per gene between the two experiments.

Nutrient availability data. The formulation of the media used in cell cul-

turing was used as Nutrient availability data. The standard Dulbecco's Modified Eagle Medium (DMEM) with different concentration of glucose and glutamine was used (see Table 7.1).

Table 7.1: NA data used in experiments in terms of their varying glucose and glutamine concentrations. The NA further contained all components like in standard DMEM medium. The concentrations were verified via HPLC measurements.

NA	Glucose [mM]	Glutamine [mM]
NA1 (Glc25Gln4)	25.0	4.0
NA2 (Glc25Gln0)	25.0	0.0
NA3 (Glc5.6Gln4)	5.6	4.0
NA4 (Glc5.6Gln0)	5.6	0.0
NA5 (Glc0Gln4)	0.0	4.0
NA6 (Glc25Gln4)	0.0	0.0

Preparing the model. During preliminary calculations, we observed that with the default Recon3D model and constraints applied in this study, biomass synthesis was not possible. We traced the problem back to the triglyceride synthesis pathway where in the default Recon3D version a 'source' reaction for triglycerides is present (which allows influx into the cell independent of the presence of triglycerides in the medium; it may be noted that in Recon3D such a source reaction is called a sink reaction, by virtue of sign notation; uptake fluxes are positive, effluxes are positive [54]). Indeed, when temporarily reactivating the triglyceride source reaction (or equivalently activating the triglyceride exchange reaction and adding triglyceride to the medium) a biomass synthesis became possible in all conditions.

Biochemically, triglycerides are synthesized starting from glycerol-3-phosphate and various lipid tails esterified to CoA. Each of these lipid tails can assume any of the three positions in the triglyceride molecule. We observed that without adding triglyceride uptake (or a ditto source) to the metabolic map, the first intermediate in the pathway (lysophosphatidic acid; the monoglyceride with a phosphate on the 3 position) could not be net-produced (see the network diagram in Figures S1 and S2). Inspection of the network revealed that this was due to the need for net input of 'Rtotal', 'Rtotal2' and 'Rtotal3' groups in this pathway for which there exists no synthesis reaction in Recon3D. This problem affects biomass flux because multiple metabolites downstream of the triglyceride synthesis pathway (e.g., Phosphatidylcholine, Phosphatidylserine, and Phosphatidylethanolamine, see Table S2), or in branches of it, are explicit components of the biomass used in Recon3D. Recon3D does have the potential to make Stearoyl-CoA, Palmityl-CoA, Oleoyl-CoA and Octadecadienoyl-CoA, but lacks the reactions to associate these with the glycerol moiety: it instead associates Rtotal, Rtotal2 and Rtotal3 to the glycerol moiety, the numbers referring to the position they take in the resulting triglyceride molecule. Recon3D worked around the ensuing problem of lack of biomass synthesis by adding a source for triglycerides. We removed such dei ex machina by forbidding source reactions

and thereby came across this problem. We solved it by equating Rtotal, Rtotal2 and Rtotal3 species in Recon3D to a single pool Rtotal and by adding a pooling reaction for lipid tails:

$$1\,\text{Stearoyl-CoA} + 1\,\text{Palmityl-CoA} + 1\,\text{Oleoyl-CoA} + 1\,\text{Octadecadienoyl-CoA} \leftrightarrow 4\,\text{Rtotal-CoA}$$

This apparent synthesis reaction merely reflects that the four acyl groups mentioned may be considered a single pool, in the sense that they can largely substitute for each other as lipid tails in biomass. With these two adjustments we have essentially reverted the Recon3D model back to how these reactions were annotated in Recon 2.2 [62, 64] but with a different Rtotal synthesis reaction. In Recon2.2 Rtotal synthesis was only possible from Palmityl-CoA and Palmitoleyl-CoA in separate reactions. From Icosanoyl-CoA and Stearoyl-CoA Rtotal2CoA and Rtotal3CoA, respectively, could be synthesized but these metabolites were not connected to anything else in the network. Our grouping of the synthesis of all Rtotal into a single reaction, has the advantage that a ratio may be imposed when this is known experimentally. We here took this ratio to be 1:1:1:1 This workaround allowed Recon3D to sustain a positive biomass flux on the medium discussed above. An alternative in which all Rtotal-CoA could be synthesized from any of the four above CoA esters was not here entertained. The resulting 'patched' version of Recon3D is available in the supplementary code and model archive and a list of its reactions and metabolites can be seen in Supplementary Excel Table S2.

Conversion RNA-seq to RAS. First, we converted the entrez gene identifiers in Recon3D to their gene symbols (e.g., 8639.1 was converted to AOC3), using the mygene module in Python (`https://pypi.org/project/mygene/`), in order to match them to genes represented in the transcriptomics datasets. Out of the 2248 genes in Recon3D, we were unable to match 103 to the dataset on the basis of their gene symbol alone. We then additionally searched the dataset for known gene aliases and this led us to identifying an additional 87 transcripts. In total, 16 genes of Recon3D (see Table S1) we could not identify. These included the three mitochondrial genes encoding the three subunits of cytochrome c oxidase. We artificially assigned all these 16 genes a TPM score equal to the maximum TPM score of the genes we could identify for each cell line. Since we do not know whether or not these genes are expressed, we did not want to artificially block them in our model analyses. Setting them to the maximal observed TPM value guarantees that they will not be limiting in any of our analyses.

The existing RNA-sequencing methodology suffers from so-called zero-inflation [65], i.e. the lack of transcripts for genes that are in fact expressed. For data integration this is problematic since a single zero may block an entire pathway. For our dataset we did indeed observe this problem. For example, in the RNA-seq data the serinepalmitoyltransferase-long-chain-base-subunit-3 gene (SPTLC3) which catalyzes the reaction synthesizing 3-dehydrosphinganine (SERPT), has a TPM of zero, and blocking this reaction (after imposing our

changes to Recon3D as discussed above) singlehandedly prevents biomass synthesis. To bypass this problem, we used a microarray dataset and calculated the ratio R of the microarray intensity divided by the genome-wide median for each gene that came associated with a TPM score of zero in the RNA-seq dataset. Then we updated such genes' TPM scores and set them equal to the calculated ratio R for the microarray dataset multiplied by the genome-wide median in the RNA-seq dataset. Below we will therefore refer to this set of zero-adjusted TPM scores as TPM*. Because the microarray dataset had scores for all genes, this removed all zeros from the dataset. We here neglected any transcriptome difference between the cell lines and experimental conditions used for the microarray and RNA-seq experiments and we assumed that the microarray data were quantitative also at low gene expression.

Recon3D contains an annotated gene-reaction coupling rule for each reaction. Using AND and OR logic this rule specifies which genes encode proteins that may help catalyze that reaction. The AND logic may be used to indicate proteins that consist of more than one subunit, or protein complexes that catalyze reactions whereas the OR logic may be used to indicate isoenzymes or alternative configurations of the protein complexes. When integrating the transcriptomics data into the map we turned these Boolean gene-reaction coupling rules into quantitative rules. Here we were inspired by MaREA methodology [31] and the E-flux approach [28]. MaREA had been developed to compare reaction activities between patients rather than between cell lines by assigning a quantitative so-called RAS to each reaction based on gene expression levels for all proteins that might be involved in the catalysis of that reaction also turning the above Boolean rules into quantitative activities. In order to do this, one needs to know the levels of the corresponding proteins and protein subunits in the cell of interest. We used the mRNA levels as a first approximation to the corresponding protein levels, assuming that the two were proportional. We assigned to each reaction a RAS by summing over isoenzymes (for OR logic) and taking minima of subunits of a complex (for AND logic) the TPM scores for the genes coupled to each reaction. In this way, isoenzymes are thought to contribute additively to the activity of a reaction whereas lack of even one subunit of an enzyme complex can linearly bring down a reaction's activity [31, 56].

Conversion NA to MUR. Once a network reconstruction is converted to a mathematical format, the inputs to the system must be defined by adding consideration of the extracellular environment. In a metabolic network like RECON3D, compounds enter and exit the extracellular environment via exchange reactions. Exchange refers to the net consumption or production of a metabolite and occurs between the extracellular compartment and the outside. In our GENSI approach the GEM is able to import only compounds that are available for cells. We introduced the Maximum Uptake Rate value as the rate of the maximum possible uptake over the time for each substance available to the model. We define a MUR_j for each exchange reaction j based on NA data for each condition as the absolute value of the difference in concentration of the substate s_{ex} in the extracellular

environment over the time t as follows:

$$MUR_j = \frac{|d[S_j]|}{dt} \approx \frac{|\Delta S_j|}{\Delta t} = \frac{|[S_{j1}] - [S_{j0}]|}{t_1 - t_0} \tag{7.1}$$

where S_j denotes the concentration of the extracellular substance j at timepoint 1 and 0. MUR has the units of mol/L/h. In this study, S_{j1} is zero due to the assumption that the full amounts of each substance could be taken up.

In the next step MUR_j is converted to a lower bound l_j of exchange reaction r_{ex} as follows:

$$l_j \leq v_j \leq u_j \tag{7.2}$$
$$l_j = -1 \times MUR_j \tag{7.3}$$

where v_j is the flux through exchange reaction j, l_j and u_j are lower and upper bounds on this flux, respectively, MUR_j is a Maximum Uptake Rate that is defined as the rate of the maximum possible uptake over time for substance j that is available to the model.

If the concentration of a substrate supplied to the cells is zero, then the bound (=upper limit) on inward exchange flux is zero ($l_j = 0$). This does not mean that the transport reaction of that metabolite is blocked (has bounds of zero) but it does mean that that the transport flux will be zero due to lack of substrate. If that extracellular concentration were to go up (by increasing the exchange reaction bound), the metabolite could be imported and transport flux would be possible.

We blocked all other uptake of metabolites so that only those compounds listed in Table S7.4 were allowed to be exchanged in. We did not alter the Recon3D default choices of which metabolites may be net produced by the in silico cell. This left 1559 metabolites which are allowed to be net-produced by the cell (Supplementary Excel Table S1). Recon3D also contains various so-called sink and demand reactions which serve as sources and sinks for certain metabolites allowing them to bypass the regular mass-balance. We blocked all such reactions.

We defined NA based on composition on the media used in experiments, we verified the concentration by HPLC after adding serum. All uptake reactions and their maximal uptake rates were taken the same across the six conditions with the exceptions of glutamine and glucose which were varied as specified in Table 7.1 and Table S7.4 in Supplementary Material.

Simulations

In all simulations in this work, we apply the computational technique of FBA [54] and FVA [61] to the human genome-wide metabolic map Recon3D [55] using the COBRA toolbox in MATLAB and Python [66, 67]. FBA entails the following linear programming problem:

$$\begin{aligned} &Z = c^\top V, \text{ such that} \\ &SV = 0V \\ &\alpha \leq V \leq \beta \end{aligned} \tag{7.4}$$

where S is the stoichiometry matrix indicating how many molecules of each metabolite are produced or consumed in each reaction, V is the vector of fluxes through all reactions including exchange reactions with the environment of the system considered, α and β are the vectors of lower and upper bounds on these fluxes, and c is a vector of weights generating the linear combination of fluxes that constitutes the objective function Z. A flux distribution resulting from FBA therefore satisfies the requirements that each metabolite is produced at the same rate as it is consumed, that the flux boundaries are not exceeded and that the flux distribution maximizes (or minimizes) the objective function Z.

Whereas FVA:

$$\begin{aligned} &\max_V / \min_V \ V_i \\ &\quad SV = 0 \\ &\quad w^\top V \geq \gamma Z_0 \\ &\quad \alpha \leq V \leq \beta \end{aligned} \tag{7.5}$$

where $Z_0 = w^\top V_0$ is the optimal solution to the FBA problem with biomass reaction as the objective function, $w^\top$ represents the biomass objective (vector of weights generating the linear combination of fluxes that constitutes the objective function Z), V is the vector of fluxes through all reactions, γ is a parameter that controls whether the analysis is done w.r.t. suboptimal network states ($0 \leq \gamma < 1$) or to the optimal state ($\gamma = 1$).

Data and source code availability All Python and MATLAB code, Jupyter Notebooks and raw data files are available as part of a Github repository `https://github.com/ThierryMondeel/HCC_flux_balance_analysis`.

In vitro experiments

Cell lines For our study we chose two HCC cell lines, i.e. the hepatitis infection-negative Huh7 and the hepatitis infection-positive PLC/PRF/5. Huh7 is a well differentiated hepatocyte-derived cellular carcinoma cell line originating in a liver tumor in a 57-year-old Japanese male. Huh7 is an immortal cell line of epithelial-like, tumorigenic cells. Its cells adhere to the surface of flasks or plates and typically grow as 2D monolayers. Furthermore, known as the Alexander hepatoma cell line, PLC/PRF/5 is a human hepatoma originally taken from the liver of a patient with primary liver cell carcinoma who was persistently infected with hepatitis B virus. Furthermore, these cells adhere to the surface of flasks or plates and grow as 2D monolayers [68, 69]. Human HCC cells (Huh-7D12 (ECACC 01042712) and PLC/PRF/5 (ECACC 85061113) were purchased from the European Collection of Authenticated Cell Cultures (Salisbury, UK).

Cell culture The Huh7 and PLC/PRF/5 cell lines were cultured in Dulbecco's modified Eagle's media (DMEM; GibcoBRL, Grand Island, NY, USA) at various initial concentrations of glucose and glutamine (M1-M6, Table 7.1), supplemented with 10% fetal bovine serum (GibcoBRL, Grand Island, NY, USA),

and a solution of 100 U/mL penicillin and 100 μg/mL streptomycin (GibcoBRL, Grand Island, NY, USA), and grown at 37 °C in a 5% CO_2 incubator at physiological pH. Cells were seeded on 25 cm^2 cell culture T flasks (Sarstedt, Numbrecht, Germany) and sub-cultured by trypsin-EDTA (GibcoBRL, Grand Island, NY, USA) treatment. During the experiments reported below the medium was not refreshed.

Growth rate/Cell proliferation assay Proliferation assays were conducted in 25 cm^2 T flasks, starting with a cell density of 8.2×105 and 5.2×105 cells/ml for Huh7 and PLC, respectively. At the time points indicated, media were collected, cells were washed and trypsinized with a 0.25% (W/V) solution of trypsin (GibcoBRL, Grand Island, NY, USA). The total number of cells in the consequent supernatant was determined by hemocytometer counting (viable plus non-viable). Mean growth rate was determined by counting six non-overlapping sets of sixteen corner squares selected at random, and these four times at each time point.

Metabolic assays The concentrations of glucose, glutamine and lactate in samples from the cells' supernatant were determined by High Performance Liquid Chromatography (HPLC) based on calibration curves made with standard solutions. The samples were taken from supernatant at the end of experiment, filtered using 0.22 μm syringe filters (BGB Analytik Vertrieb GmbH, Rheinfelden, Germany) and stored until measurement in -80 °C. HPLC was performed using the HPLC-DAD RID LC-20AT Prominence (Shimadzu, DC, USA) machine with a UV Diode Array Detector SPD-M30A NexeraX2 or/and a Refractive Index Detector RID 20A and an analytical ion-exclusion Rezex ROA-Organic Acid H+(8%) column (250×4.6 mm) with guard column (Phenomenex, Torrance, CA, USA) (5 mM H2SO4 in MilliQ water (18.2 MΩ), isocratic, 0.15 mL/min. flow rate). Injection volume was 15 μL (Autosampler: SIL-20AC, Prominence, Shimadzu), column oven temperature was 55 °C (Column oven: CTO-20A, Prominence, Shimadzu) and the pressure was 29 bar.

Supplementary Information

Supplementary Files

All Python and MATLAB code, Jupyter Notebooks, Supplementary Tables and raw data files are available as part of a Github repository at https://github.com/ThierryMondeel/HCC_flux_balance_analysis and through the online publication at https://doi.org/10.3390/biom11040490.

Supplementary Figures

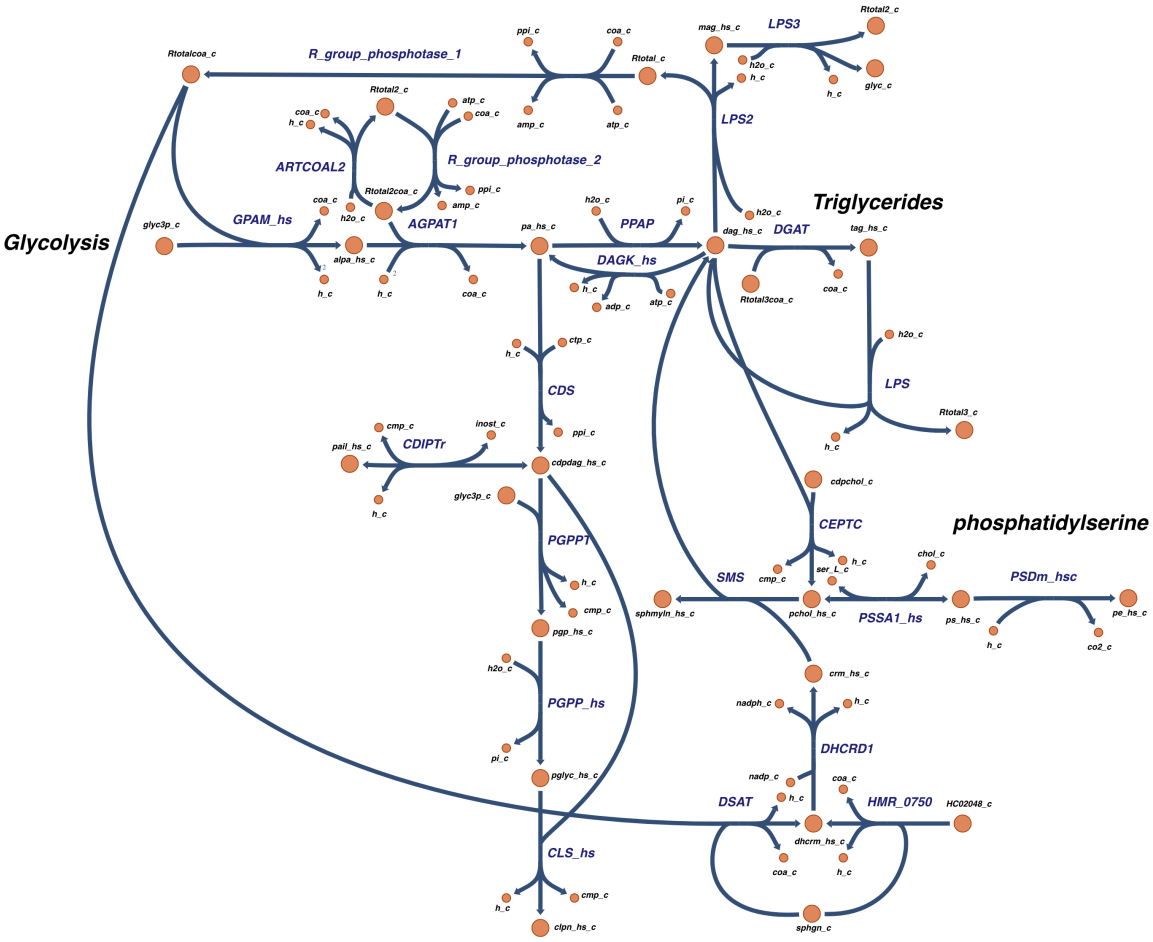


Figure S7.1: Network diagram of the subnetwork of the published version of Recon3.01 related to lipid synthesis from glycerol-3-phosphate (referred to as *glyc3p_c*) drawn using Escher [70]. See Fig. S7.2 for the updated network diagram.

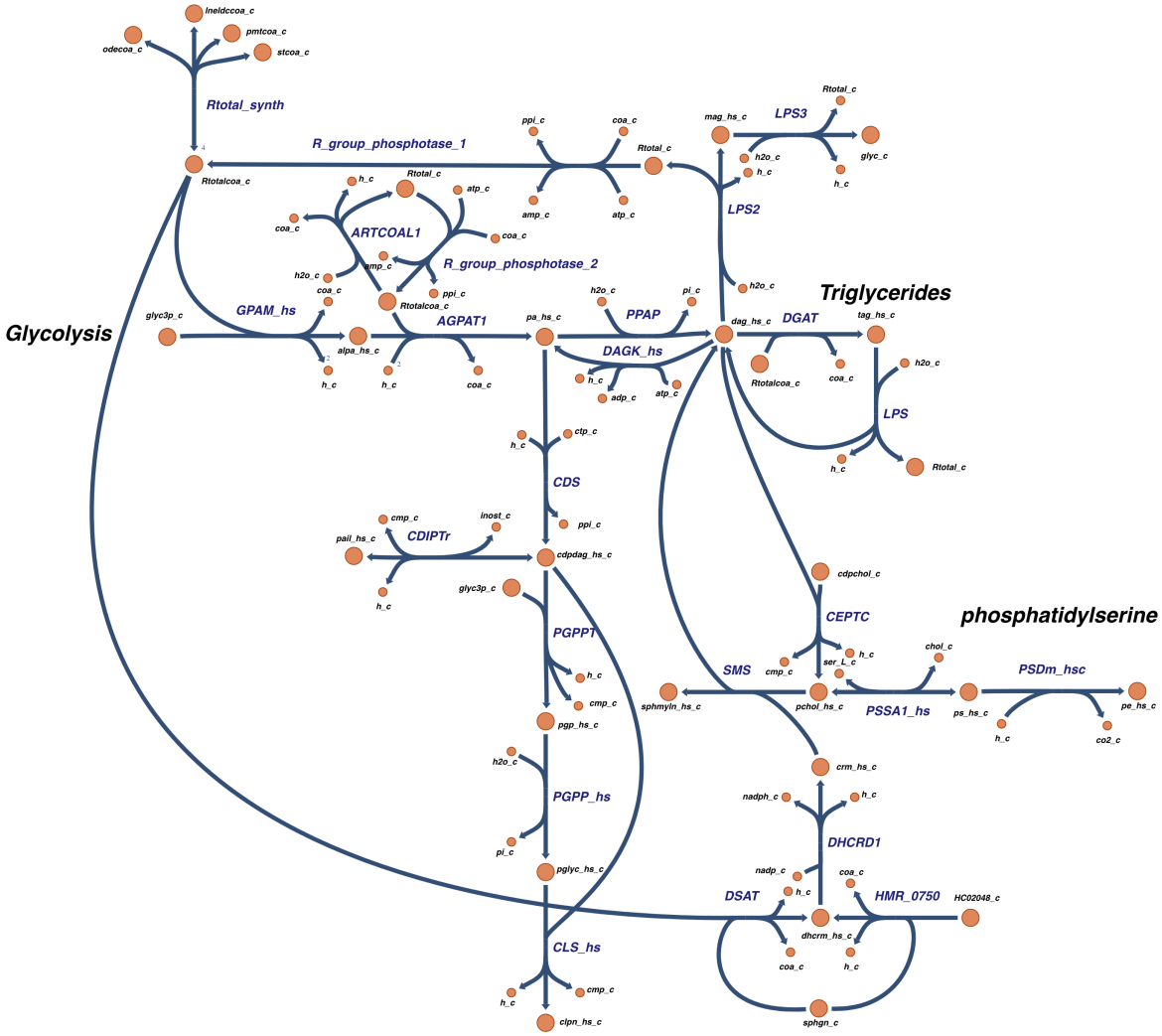


Figure S7.2: Network diagram of the updated subnetwork in Recon3.01 related to lipid synthesis from glycerol-3-phosphate (referred to as *glyc3p_c*) drawn using Escher [70]. We enhanced the pathway by adding a synthesis reaction for Rtotal and by equating the various forms of Rtotal by adding reversible reactions between them (see main text).

Entrez ID	Symbol	Alias	Name
201288.1	NOS2P2	NOS2B	nitric oxide synthase 2 pseudogene 2
645740.1	NOS2P1	NOS2C	nitric oxide synthase 2 pseudogene 1
728441.1	GGT2	GGT, GGT 2	gamma-glutamyltransferase 2
102724560.1	CBSL	CBS	cystathionine-beta-synthase like
65263.1	PYCR3	PYCRL	pyrroline-5-carboxylate reductase 3
8781.1	PSPHP1	CO9, PSPHL	phosphoserine phosphatase pseudogene 1
100507855.1	–	–	–
644378.1	GCNT2P1	GCNT2P, GCNT6	GCNT2 pseudogene 1
284004.1	HEXD	HEXDC	hexosaminidase D
8041.1	–	–	–
100288072.1	SDR42E2		short chain dehydrogenase/reductase family 42E, member 2
340811.1	AKR1C8P	AKR1CL1	aldo-keto reductase family 1 member C8, pseudogene
0	–	–	–
4514.1	COX3	COIII, MTCO3	cytochrome c oxidase III
4513.1	COX2	COII, MTCO2	cytochrome c oxidase subunit II
4512.1	COX1	COI, MTCO1	cytochrome c oxidase subunit I

Table S7.1: List of Entrez identifiers for genes present in Recon3D that we were not able to identify in the RNA-seq dataset by Ma et al. [57] after converting the Entrez ID to the respective gene symbol and possible aliases. Red cells concern Entrez identifiers that do not exist, whereas orange cells concern Entrez identifiers that have been withdrawn from NCBI. The 16 genes in white or orange cells were assumed to be expressed at excess levels.

Metabolite	Coeff.	Metabolite	Coeff.
Adenosine Triphosphate	-20.7045	Deoxyadenosine Triphosphate	-0.01318
Adenosine Diphosphate	20.65082	Deoxycytidine-5′-Triphosphate	-0.00944
Orthophosphate	20.65082	Deoxythymidine-5′-Triphosphate	-0.01309
Sphingomyelin	-0.01749	Deoxyguanosine-5′-Triphosphate	-0.0099
Water	-20.6508	Guanosine-5′-Triphosphate	-0.03612
Proton	20.6508	Cytidine-5′-Triphosphate	-0.03904
L-Alanine	-0.50563	1-Phosphatidyl-1D-Myo-Inositol	-0.02332
L-Arginine	-0.35926	Phosphatidylethanolamine	-0.05537
L-Asparagine	-0.27943	D-Glucose 6-Phosphate	-0.27519
L-Aspartate	-0.35261	Phosphatidylcholine	-0.15446
Cholesterol	-0.0204	Phosphatidylglycerol	-0.00291
Cardiolipin	-0.01166	Phosphatidylserine	-0.00583
L-Cysteine	-0.04657	Uridine-5′-TrIphosphate	-0.05345
L-Glutamine	-0.326	L-Glutamate	-0.38587
Glycine	-0.53889	L-Histidine	-0.12641
L-Isoleucine	-0.28608	L-Leucine	-0.54554
L-Lysine	-0.59211	L-Methionine	-0.15302
L-Phenylalanine	-0.25947	L-Tyrosine	-0.15967
L-Serine	-0.39253	L-Proline	-0.41248
L-Tryptophan	-0.01331	L-Threonine	-0.31269
L-Valine	-0.35261		

Table S7.2: Metabolic components of the biomass reaction in Recon3D listed along with the stoichiometry at which they were taken to contribute to the biomass pseudo reaction. The stoichiometric coefficient is negative if the compound was taken to be consumed in the biomass synthesis reaction and positive if it was produced therein. It was obtained [55] by determining the mass-fractional contributions of each biomass precursor to the dry weight of the biomass (i.e. in terms of milligram precursor per gram of dry weight (DW)). This fraction is then divided by the molecular weight of that precursor to obtain the above coefficient, which then is in unit mmol precursor per gram of biomass dry weight (g DW). The biomass reaction equation is then formulated as the sum of each metabolite precursor multiplied by its above coefficient. The biomass reaction rate is the specific growth rate and has units (g DW / g DW)/h. The specific consumption rates of the metabolites are obtained by multiplying their above coefficients by the specific growth rate and thereby are in units mmol/(g DW)/h [71]. Multiplying these specific consumption rates by the biomass concentration in g DW/dm^3 of the culture vessel, one obtains the consumption flux in terms of mM/h.

ID	Name	Reaction
ATPS4mi	ATP synthase	adp_m + 4.0 h_i + pi_m → atp_m + h2o_m + 3.0 h_m
NADH2_u10mi	NADH dehydrogenase	5.0 h_m + nadh_m + q10_m → 4.0 h_i + nad_m + q10h2_m
CYOR_u10mi	ubiquinol-cytochrome c reductase	2.0 ficytC_m + 2.0 h_m + q10h2_m → 2.0 focytC_m + 4.0 h_i + q10_m
CYOOm2i	cytochrome c oxidase	4.0 focytC_m + 8.0 h_m + o2_m → 4.0 ficytC_m + 2.0 h2o_m + 4.0 h_i
CYOOm3i	cytochrome c oxidase	4.0 focytC_m + 7.92 h_m + o2_m → 4.0 ficytC_m + 1.96 h2o_m + 4.0 h_i + 0.02 o2s_m
PDHm	pyruvate dehydrogenase	coa_m + nad_m + pyr_m → accoa_m + co2_m + nadh_m

Table S7.3: List of the oxidative phosphorylation reactions considered in Fig. 7.5 by their reaction ID in Recon3D and with their reaction as specified in Recon3D.

Metabolite	Uptake bound	Metabolite	Uptake bound	Metabolite	Uptake bound
Arginine	0.40	Bicarbonate	44	Nicotinamide	0.033
Choline	0.028	Histidine	0.2	Valine	0.80
Cysteine	0.20	Isoleucine	0.80	Phenylalanine	0.4
Iron (Fe3+)	0.00024	Myo-Inositol	0.04	Phosphate	0.9
Folate	0.0091	Potassium	5.3	(R)-Pantothenate	0.0083
Glycine	0.40	Methionine	0.20	Serine	0.40
Tyrosine	0.40	Sodium	155	Sulfate	0.81
Glucose	[0, 5.6, 25]	Leucine	0.80	Pyridoxine	0.019
Glutamine	[0, 4.0]	Lysine	0.80	Riboflavin	0.0011
Thiamin	0.011	Threonine	0.80	Tryptophan	0.078

Table S7.4: Concentrations of metabolites in the DMEM media in mM as listed in the manufacturer's formulation. Our media M1-M6 only differ in the glucose and glutamine concentrations (colored in green) and were exempt of ammonia. All concentrations were effected in-silico as maximal uptake rates which shape and constrain the flux cone of the solutions in FBA (see text). For water, carbon dioxide and oxygen, a not limiting uptake bound of 1000 was taken. Oxygen was considered non-limiting because the concentration in the medium was in the order of 0.2-0.4 mM (corresponding to air saturated saline) [72] whereas the K_m of cytochrome oxidase for oxygen is some 0.01 mM [73]. Carbon dioxide was considered non-limiting due to its continuous replenishment in the medium.

References

[1] M. Planck. *Scientific Autobiography and Other Papers*. New York, NY: F. Gaynor, 1949.

[2] R. J. DeBerardinis and N. S. Chandel. "We need to talk about the Warburg effect". *Nature Metabolism* 2 (2020), pp. 127–129. 10.1038/s42255-020-0172-2.

[3] M. V. Liberti and J. W. Locasale. "The Warburg Effect: How Does it Benefit Cancer Cells?" *Trends in Biochemical Sciences* 41 (2016), pp. 211–218. 10.1016/j.tibs.2015.12.001.

[4] R. J. DeBerardinis and N. S. Chandel. "Fundamentals of cancer metabolism". *Science Advances* 2 (2016), e1600200. 10.1126/sciadv.1600200.

[5] N. N. Pavlova and C. B. Thompson. "The Emerging Hallmarks of Cancer Metabolism". *Cell Metabolism* 23 (2016), pp. 27–47. 10.1016/j.cmet.2015.12.006.

[6] C. Corbet and O. Feron. "Metabolic and mind shifts". *Current Opinion in Clinical Nutrition and Metabolic Care* 18 (2015), pp. 346–353. 10.1097/MCO.0000000000000178.

[7] O. Warburg, K. Posener, and E. Negelein. "Über den Stoffwechsel der Tumoren Biochemische". *Zeitschrift* 152 (1923), pp. 319–344.

[8] E. Racker. "Bioenergetics and the problem of tumor growth." *American scientist* 60 (1972), pp. 56–63.

[9] A. M. Abdel-Haleem *et al.* "The Emerging Facets of Non-Cancerous Warburg Effect". *Frontiers in Endocrinology* 8 (2017). 10.3389/fendo.2017.00279.

[10] I. San-Millán and G. A. Brooks. "Reexamining cancer metabolism: lactate production for carcinogenesis could be the purpose and explanation of the Warburg Effect". *Carcinogenesis* (2016), bgw127. 10.1093/carcin/bgw127.

[11] K. Kitamura *et al.* "Proliferative activity in hepatocellular carcinoma is closely correlated with glucose metabolism but not angiogenesis". *Journal of Hepatology* 55 (2011), pp. 846–857. 10.1016/j.jhep.2011.01.038.

[12] V. Iansante *et al.* "PARP14 promotes the Warburg effect in hepatocellular carcinoma by inhibiting JNK1-dependent PKM2 phosphorylation and activation". *Nature Communications* 6 (2015), p. 7882. 10.1038/ncomms8882.

[13] A. E. Vaughn and M. Deshmukh. "Glucose metabolism inhibits apoptosis in neurons and cancer cells by redox inactivation of cytochrome c". *Nature Cell Biology* 10 (2008), pp. 1477–1483. 10.1038/ncb1807.

[14] D. Y. Gui, C. A. Lewis, and M. G. Vander Heiden. "Allosteric Regulation of PKM2 Allows Cellular Adaptation to Different Physiological States". *Science Signaling* 6 (2013), pe7–pe7. 10.1126/scisignal.2003925.

[15] S. Romero-Garcia, M. M. B. Moreno-Altamirano, H. Prado-Garcia, and F. J. Sánchez-García. "Lactate Contribution to the Tumor Microenvironment: Mechanisms, Effects on Immune Cells and Therapeutic Relevance". *Frontiers in Immunology* 7 (2016). 10.3389/fimmu.2016.00052.

[16] M. G. Vander Heiden and R. J. DeBerardinis. "Understanding the Intersections between Metabolism and Cancer Biology". *Cell* 168 (2017), pp. 657–669. 10.1016/j.cell.2016.12.039.

[17] R. J. DeBerardinis *et al.* "Beyond aerobic glycolysis: Transformed cells can engage in glutamine metabolism that exceeds the requirement for protein and nucleotide synthesis". *Proceedings of the National Academy of Sciences* 104 (2007), pp. 19345–19350. 10.1073/pnas.0709747104.

[18] B. J. Altman, Z. E. Stine, and C. V. Dang. "From Krebs to clinic: glutamine metabolism to cancer therapy". *Nature Reviews Cancer* 16 (2016), pp. 619–634. 10.1038/nrc.2016.71.

[19] A. A. Cluntun, M. J. Lukey, R. A. Cerione, and J. W. Locasale. "Glutamine Metabolism in Cancer: Understanding the Heterogeneity". *Trends in Cancer* 3 (2017), pp. 169–180. 10.1016/j.trecan.2017.01.005.

[20] C. T. Hensley, A. T. Wasti, and R. J. DeBerardinis. "Glutamine and cancer: cell biology, physiology, and clinical opportunities". *Journal of Clinical Investigation* 123 (2013), pp. 3678–3684. 10.1172/JCI69600.

[21] D. R. Wise and C. B. Thompson. "Glutamine addiction: a new therapeutic target in cancer". *Trends in Biochemical Sciences* 35 (2010), pp. 427–433. 10.1016/j.tibs.2010.05.003.

[22] C. Damiani *et al.* "A metabolic core model elucidates how enhanced utilization of glucose and glutamine, with enhanced glutamine-dependent lactate production, promotes cancer cell growth: The WarburQ effect". *PLOS Computational Biology* 13 (2017). Ed. by D. A. Beard, e1005758. 10.1371/journal.pcbi.1005758.

[23] C. Gu *et al.* "Current status and applications of genome-scale metabolic models". *Genome Biology* 20 (2019), p. 121. 10.1186/s13059-019-1730-3.

[24] D. J. Cook and J. Nielsen. "Genome-scale metabolic models applied to human health and disease". *Wiley Interdisciplinary Reviews: Systems Biology and Medicine* 9 (2017), e1393. 10.1002/wsbm.1393.

[25] M. Tian and J. L. Reed. "Integrating Proteomic or Transcriptomic Data into Metabolic Models Using Linear Bound Flux Balance Analysis". *Bioinformatics* (2018). 10.1093/bioinformatics/bty445.

[26] A. Richelle, C. Joshi, and N. E. Lewis. "Assessing key decisions for transcriptomic data integration in biochemical networks". *PLOS Computational Biology* 15 (2019). Ed. by V. Hatzimanikatis, e1007185. 10.1371/journal.pcbi.1007185.

[27] M. Åkesson, J. Förster, and J. Nielsen. "Integration of gene expression data into genome-scale metabolic models". *Metabolic Engineering* 6 (2004), pp. 285–293. 10.1016/j.ymben.2003.12.002.

[28] C. Colijn *et al.* "Interpreting Expression Data with Metabolic Flux Models: Predicting Mycobacterium tuberculosis Mycolic Acid Production". *PLoS Computational Biology* 5 (2009). Ed. by J. A. Papin, e1000489. 10.1371/journal.pcbi.1000489.

[29] S. A. Becker and B. O. Palsson. "Context-Specific Metabolic Networks Are Consistent with Experiments". *PLoS Computational Biology* 4 (2008). Ed. by H. M. Sauro, e1000082. 10.1371/journal.pcbi.1000082.

[30] T. Shlomi *et al.* "Network-based prediction of human tissue-specific metabolism". *Nature Biotechnology* 26 (2008), pp. 1003–1010. 10.1038/nbt.1487.

[31] A. Graudenzi *et al.* "Integration of transcriptomic data and metabolic networks in cancer samples reveals highly significant prognostic power". *Journal of Biomedical Informatics* 87 (2018), pp. 37–49. 10.1016/j.jbi.2018.09.010.

[32] N. Vlassis, M. P. Pacheco, and T. Sauter. "Fast Reconstruction of Compact Context-Specific Metabolic Network Models". *PLoS Computational Biology* 10 (2014). Ed. by C. A. Ouzounis, e1003424. 10.1371/journal.pcbi.1003424.

[33] R. Agren *et al.* "Identification of anticancer drugs for hepatocellular carcinoma through personalized genome-scale metabolic modeling". *Molecular Systems Biology* 10 (2014), p. 721. 10.1002/msb.145122.

[34] D. R. Hyduke, N. E. Lewis, and B. Ø. Palsson. "Analysis of omics data with genome-scale models of metabolism". *Mol. BioSyst.* 9 (2013), pp. 167–174. 10.1039/C2MB25453K.

[35] S. Opdam *et al.* "A Systematic Evaluation of Methods for Tailoring Genome-Scale Metabolic Models". *Cell Systems* 4 (2017), 318–329.e6. 10.1016/j.cels.2017.01.010.

[36] T. Pfau, M. P. Pacheco, and T. Sauter. "Towards improved genome-scale metabolic network reconstructions: unification, transcript specificity and beyond". *Briefings in Bioinformatics* (2015), bbv100. 10.1093/bib/bbv100.

[37] A. Schultz and A. A. Qutub. "Reconstruction of Tissue-Specific Metabolic Networks Using CORDA". *PLOS Computational Biology* 12 (2016). Ed. by C. D. Maranas, e1004808. 10.1371/journal.pcbi.1004808.

[38] E. C. Lien and M. G. Vander Heiden. "A framework for examining how diet impacts tumour metabolism". *Nature Reviews Cancer* 19 (2019), pp. 651–661. 10.1038/s41568-019-0198-5.

[39] M. R. Sullivan *et al.* "Quantification of microenvironmental metabolites in murine cancers reveals determinants of tumor nutrient availability". *eLife* 8 (2019). 10.7554/eLife.44235.

[40] A. Muir, L. V. Danai, and M. G. Vander Heiden. "Microenvironmental regulation of cancer cell metabolism: implications for experimental design and translational studies". *Disease Models & Mechanisms* 11 (2018). 10.1242/dmm.035758.

[41] A. Muir and M. G. Vander Heiden. "The nutrient environment affects therapy". *Science* 360 (2018), pp. 962–963. 10.1126/science.aar5986.

[42] M. P. Joanna Wietrzyk. "Cancer – Could it be Cured? A Spontaneous Regression of Cancer, Cancer Energy Metabolism, Hyperglycemia-Hypoglycemia, Metformin, Warburg and Crabtree Effects and a New Perspective in Cancer Treatment". *Journal of Cancer Science & Therapy* 06 (2014). 10.4172/1948-5956.1000249.

[43] A. Luengo, D. Y. Gui, and M. G. Vander Heiden. "Targeting Metabolism for Cancer Therapy". *Cell Chemical Biology* 24 (2017), pp. 1161–1180. 10.1016/j.chembiol.2017.08.028.

[44] S. M. Davidson *et al.* "Environment Impacts the Metabolic Dependencies of Ras-Driven Non-Small Cell Lung Cancer". *Cell Metabolism* 23 (2016), pp. 517–528. 10.1016/j.cmet.2016.01.007.

[45] D. Y. Gui *et al.* "Environment Dictates Dependence on Mitochondrial Complex I for NAD+ and Aspartate Production and Determines Cancer Cell Sensitivity to Metformin". *Cell Metabolism* 24 (2016), pp. 716–727. 10.1016/j.cmet.2016.09.006.

[46] J. R. Mayers *et al.* "Tissue of origin dictates branched-chain amino acid metabolism in mutant Kras -driven cancers". *Science* 353 (2016), pp. 1161–1165. 10.1126/science.aaf5171.

[47] J. R. Cantor *et al.* "Physiologic Medium Rewires Cellular Metabolism and Reveals Uric Acid as an Endogenous Inhibitor of UMP Synthase". *Cell* 169 (2017), 258–272.e17. 10.1016/j.cell.2017.03.023.

[48] J. Vande Voorde *et al.* "Improving the metabolic fidelity of cancer models with a physiological cell culture medium". *Science Advances* 5 (2019). 10.1126/sciadv.aau7314.

[49] C. Corbet and O. Feron. "Tumour acidosis: from the passenger to the driver's seat". *Nature Reviews Cancer* 17 (2017), pp. 577–593. 10.1038/nrc.2017.77.

[50] E. Persi *et al.* "Systems analysis of intracellular pH vulnerabilities for cancer therapy". *Nature Communications* 9 (2018), p. 2997. 10.1038/s41467-018-05261-x.

[51] O. D. K. Maddocks *et al.* "Modulating the therapeutic response of tumours to dietary serine and glycine starvation." *Nature* 544 (2017), pp. 372–376. 10.1038/nature22056.

[52] M. Zampieri *et al.* "Regulatory mechanisms underlying coordination of amino acid and glucose catabolism in Escherichia coli". *Nature Communications* 10 (2019), p. 3354. 10.1038/s41467-019-11331-5.

[53] A. Varma and B. O. Palsson. "Metabolic Flux Balancing: Basic Concepts, Scientific and Practical Use". *Nature Biotechnology* 12 (1994), pp. 994–998. 10.1038/nbt1094-994.

[54] J. D. Orth, I. Thiele, and B. Ø. Palsson. "What is flux balance analysis?" *Nature Biotechnology* 28 (2010), pp. 245–248. 10.1038/nbt.1614.

[55] E. Brunk *et al.* "Recon3D enables a three-dimensional view of gene variation in human metabolism". *Nature Biotechnology* 36 (2018), pp. 272–281. 10.1038/nbt.4072.

[56] C. Damiani *et al.* "Integration of single-cell RNA-seq data into population models to characterize cancer metabolism". *PLOS Computational Biology* 15 (2019). Ed. by C. Kaleta, e1006733. 10.1371/journal.pcbi.1006733.

[57] M. K. F. Ma *et al.* "Stearoyl-CoA desaturase regulates sorafenib resistance via modulation of ER stress-induced differentiation". *Journal of Hepatology* 67 (2017), pp. 979–990. 10.1016/j.jhep.2017.06.015.

[58] K. N. Rybakova *et al.* "Multiplex Eukaryotic Transcription (In)activation: Timing, Bursting and Cycling of a Ratchet Clock Mechanism." *PLoS computational biology* 11 (2015), e1004236. 10.1371/journal.pcbi.1004236.

[59] B. H. ter Kuile and H. V. Westerhoff. "Transcriptome meets metabolome: hierarchical and metabolic regulation of the glycolytic pathway". *FEBS Letters* 500 (2001), pp. 169–171. 10.1016/S0014-5793(01)02613-8.

[60] K. Smallbone *et al.* "A model of yeast glycolysis based on a consistent kinetic characterisation of all its enzymes". *FEBS Letters* 587 (2013), pp. 2832–2841. 10.1016/j.febslet.2013.06.043.

[61] R. Mahadevan and C. Schilling. "The effects of alternate optimal solutions in constraint-based genome-scale metabolic models". *Metabolic Engineering* 5 (2003), pp. 264–276. 10.1016/j.ymben.2003.09.002.

[62] I. Thiele *et al.* "A community-driven global reconstruction of human metabolism". *Nature Biotechnology* 31 (2013), pp. 419–425. 10.1038/nbt.2488.

[63] Y. D. Shaul *et al.* "MERAV: a tool for comparing gene expression across human tissues and cell types". *Nucleic Acids Research* 44 (2016), pp. D560–D566. 10.1093/nar/gkv1337.

[64] N. Swainston *et al.* "Recon 2.2: from reconstruction to model of human metabolism". *Metabolomics* 12 (2016), p. 109. 10.1007/s11306-016-1051-4.

[65] S. C. Hicks, F. W. Townes, M. Teng, and R. A. Irizarry. "Missing data and technical variability in single-cell RNA-sequencing experiments". *Biostatistics* 19 (2018), pp. 562–578. 10.1093/biostatistics/kxx053.

[66] L. Heirendt *et al.* "Creation and analysis of biochemical constraint-based models using the COBRA Toolbox v.3.0". *Nature Protocols* 14 (2019), pp. 639–702. 10.1038/s41596-018-0098-2.

[67] A. Ebrahim, J. A. Lerman, B. O. Palsson, and D. R. Hyduke. "COBRApy: COnstraints-Based Reconstruction and Analysis for Python". *BMC Systems Biology* 7 (2013), p. 74. 10.1186/1752-0509-7-74.

[68] R. J. Daemer *et al.* "PLC/PRF/5 (Alexander) hepatoma cell line: further characterization and studies of infectivity." *Infection and immunity* 30 (1980), pp. 607–11.

[69] A. C. Krelle, A. S. Okoli, and G. L. Mendz. "Huh-7 Human Liver Cancer Cells: A Model System to Understand Hepatocellular Carcinoma and Therapy". *Journal of Cancer Therapy* 04 (2013), pp. 606–631. 10.4236/jct.2013.42078.

[70] Z. A. King *et al.* "Escher: A Web Application for Building, Sharing, and Embedding Data-Rich Visualizations of Biological Pathways". *PLOS Computational Biology* 11 (2015). Ed. by P. P. Gardner, e1004321. 10.1371/journal.pcbi.1004321.

[71] I. Thiele and B. Ø. Palsson. "A protocol for generating a high-quality genome-scale metabolic reconstruction". *Nature Protocols* 5 (2010), pp. 93–121. 10.1038/nprot.2009.203.

[72] T. L. Place, F. E. Domann, and A. J. Case. "Limitations of oxygen delivery to cells in culture: An underappreciated problem in basic and translational research". *Free Radical Biology and Medicine* 113 (2017), pp. 311–322. 10.1016/j.freeradbiomed.2017.10.003.

[73] K. Krab, H. Kempe, and M. Wikström. "Explaining the enigmatic KM for oxygen in cytochrome c oxidase: A kinetic model". *Biochimica et Biophysica Acta (BBA) - Bioenergetics* 1807 (2011), pp. 348–358. 10.1016/j.bbabio.2010.12.015.

CHAPTER 8

General Discussion

"But one thing is certain: to understand the whole you must look at the whole."

— Henrik Kacser [1]

Systems biology combines experimental and computational branches into a spiral of ever increasing understanding [2]. The two branches are intimately interwoven since computational systems biology critically depends on measurements, whose coverage and accuracy have a direct bearing on the value of computational models [3]. *Vice versa*, computational models often suggest novel hypotheses, which may guide further experimental testing. This interplay was present in the preceding chapters, albeit with an emphasis on the computational branch, and the results have shed light on several important biological questions, raised new questions and suggested multiple avenues for future research.

In this general discussion, I will present a bird's eye view of the previous chapters in light of this interplay between computational and experimental research, of the next steps they inspire and of how they relate to the general explosion of computing power and big data. I hope to bring these questions and future challenges together under a single umbrella.

8.1 Networks, coupling, dynamics and objectives

Ch. 2-7 dealt with seemingly separate and unique questions and challenges that we encountered (or invented) in the laboratory where I performed this research. However, every chapter in this thesis has fundamentally viewed bacterial and eukaryotic cells as a complex system, involving a network of interacting components, the behavior of which is determined by that network's interaction with the environment instead of by single components and exhibits emergent behavior.

The aims of systems biology are to discover general principles about how functional properties and behavior of living organisms arise from the interactions of their constituents, to understand these principles and systems, and to predict their future states and behavior [2, 4]. Did we make any progress towards these aims in Ch. 2-7? We discussed in Ch. 1 that it is crucial to understand complex systems in terms of (i) what the components can do and how they interact, (ii) the system's inputs and outputs and the relationships (coupling) between them, (iii) the dynamics across time of components and their interactions, and, when there is selection pressure such as in living or technological systems, (iv) the system's goal or objective, if any, and how this objective is achieved through the

interactions between the components. As a consequence, network descriptions, network analysis, media and secretion products, predictions of system dynamics, and the optimization of biological objectives featured heavily in this thesis. Below, I will discuss what we learned from Ch. 2-7 with regard to these four aspects of biological complex systems.

From network (re)wiring to functionality

Ch. 2-7 all discuss network descriptions for the systems relevant for the problem at hand and have resulted in several new insights pertaining to the network wiring, network visualization and the limitations of looking *only* at networks. In Ch. 2 the exact nature of the interactions between cell cycle regulatory components was one of the main themes. We contrasted model output for 11 different network structures where the only difference between these models lies in the addition, removal or different treatment of interactions in the base network [5]. The analysis in Ch. 2 reinforces our understanding of the importance of positive (PFL) and negative (NFL) feedback loop motifs for the occurrence of oscillations. The model output suggests in particular that the Clb3 PFL and the NFLs from Clb2 on Clb5 through the MBF transcription factor and through the APC improve the ability to obtain oscillatory model output. Furthermore, the results for Design 7 suggest that the currently hypothetical transcriptional inhibition of Clb2 and Clb3 by Sic1 should have a positive effect on oscillatory behavior of the system and suggests this interaction to be targeted for experimental validation. This is an example of computational work suggesting downstream experimental analyses.

Ch. 3 presents and analyses a dataset that improves our understanding of the network of interactions between genes and proteins in budding yeast. Through the analysis of ChIP-exo experiments for the transcription factors Fkh1 and Fkh2 we present the scientific community with updated information on their target genes. Ch. 3 suggests novel target genes, questions some previously reported targets and strengthens the evidence for other target genes of Fkh1 and Fkh2. Finally, the analysis highlights a potential pathway by which Fkh1 and Fkh2 may couple the cell cycle with metabolism by influencing a set of metabolic genes.

Ch. 4 presents GEMMER, a web-based visualization tool which improves the ability of the scientific community to query and visualize the large amount of information available for budding yeast interaction networks. GEMMER can play a role in bridging computational and experimental work by helping to place new experimental findings in their network context and by suggesting related genes and proteins that may be relevant for further analysis. As a testament to this, Ch. 3 made extensive use of the database that drives GEMMER to analyze the forkhead ChIP-exo dataset. It did this in terms of gene annotation, target gene expression windows, target gene functions and the (metabolic) pathways that the target genes play a role in.

In Ch. 5 we highlighted how organisms, and *Cl. ljungdahlii* in particular, may dynamically alter their network by subtly changing the expression of enzymes catalyzing similar reactions in order to navigate the trade-off between rate and yield. This (potentially dynamic) method of rewiring the network, even within

a single pathway, to navigate changing needs may apply to all organisms and sheds new light on the fluidity of life and the potential robustness "hidden" in the genome and proteome. We linked this idea of gear-shifting to the existence of pseudo-enzymes which may play a role as substitutes for better-understood gene-products when the situation calls for changes in pathway flux rate and yield.

In Ch. 6 we show how a biomarker prediction algorithm that reasons solely based on oversimplified network characteristics and medium composition fails to correctly predict biomarkers for glutathione-mediated drug detoxification because of the presence of a specific network characteristic: an oxoproline salvage loop. This loop leads to incorrect predictions when utilizing a model that ignores dynamics and kinetics but not when these details are included.

Ch. 7 contains a detailed accounting of transcriptomics in two hepatocellular carcinoma cell lines mapped onto the human metabolic network. This analysis shows that there are significant differences between these cell lines in the expression of enzymes that form the connections between metabolites in the metabolic network. Using flux balance analysis we showed that these differences, which may be thought of as differences in network connectivity and edge capacity, are predicted to be potentially functionally relevant.

Insights into the coupling between processes

The results from Ch. 2 can be viewed in light of the coupling between early and late cell cycle processes and in terms of the start and end of the yeast cell cycle. It is this coupling that puts the “cycle” in cell cycle. Especially the hypothetical transcriptional inhibition, directly or indirectly, of CLB2 and CLB3 by Sic1 provides a connection from the G2/mitotic cyclins back to the start of the cell cycle. If this interaction were to be validated it would link the rise of Sic1 concentrations in M phase to the abolishing of CLB2,3 transcription which promotes mitosis. Sic1 would then continue to play a role from late M phase into G1 and G1/S phase, the initiation of S phase coinciding with the disappearance of Sic1. Still, given the hypothetical interaction, this would lift the inhibition of CLB2,3 transcription bringing us back to G2/M phase.

The potential pathways, unearthed in Ch. 3, by which Fkh1 and Fkh2 may couple the cell cycle with metabolism by influencing a set of metabolic genes, exemplify how cells may organize the interconnections between their internal processes. Fkh1 and Fkh2 may provide, or tune, the uni- or bidirectional linkage long sought for [6] between the cell cycle and various metabolic processes.

GEMMER, the tool developed in Ch. 4, enables one to better than before visualize connections between processes such as cell cycle and metabolism by allowing network interaction diagrams to be drawn together with functional and temporal information. Such tools are important for understanding how intracellular and intercellular modules and systems interact and cross-talk [7].

In Ch. 5 we reviewed the concept of coupled processes in terms of thermodynamics and network rewiring. Cells need to couple processes that require Gibbs free energy input to processes that release Gibbs free energy. As shown in Fig. 5.5, at lower values of the free energy of anabolism, higher phenomenological

stoichiometries Z lead to higher anabolic fluxes, but at more challenging free energies of anabolism, the systems with lower values of Z lead to faster anabolism. We propose that a variomatic strategy could optimize the gear shifting so that always the highest anabolic flux is attained at every free energy of anabolism. We illustrated this concept by turning to acetogenic bacteria and *Cl. ljungdahlii* in particular, which use the Wood-Ljungdahl pathway to produce acetate from carbon-dioxide and hydrogen. Table 5.1 and Fig. 5.7 show that by shuffling closely-related enzymes in the WLP, optimal ATP production coupled to acetogenesis may be maintained for different Gibbs free energies of ATP hydrolysis and that this could work nearly seamlessly.

In Ch. 6 we characterized the shortcomings in a method based on flux variability analysis which aims to link metabolic perturbations to downstream shifts in metabolic concentrations in extracellular fluids or the urine. This chapter in a way carries a negative tone, the method is (overly) sensitive to some parameter settings and particular network properties (like the oxoproline loop and multiple metabolic reactions performing the same function in the case of phenylketonuria (PKU)). The positive side thereby is that the method might be extended to serve the aim of predicting downstream effects of perturbations relevant for applications of systems biology to health care. One conclusion we draw from our analysis is that in the case of metabolic biomarkers, efforts should focus on building accurate kinetic models. However, the flux variability approach has been shown to work reasonable well for inborn errors of metabolism and should work given limited assumptions on the network. We conclude that the flux variability method for biomarker prediction may continue to serve as an investigative tool, provided that it is not relied on in any more definitive sense.

Ch. 5-7 discussed input and output of biological systems in terms of cell culture media. The latter served as input (i.e. substrates) to flux balance analysis models and led to predicted model output in terms of secreted compounds. Ch. 7 presented a novel flavor of flux balance analysis which takes into account medium concentrations of metabolites and matches them with intracellular capacity constraints reflected by enzyme expression. We show how changing the medium affects growth rate, the utilization of certain pathways, and the secretion of products. Since the model predictions are in agreement with the experimental observations, this provides a counter-example to Ch. 6 where we failed to find such agreement. Here, non-kinetic models may provide useful output. Yet, also in that case, the predictions are conditional upon properties of the system that are most often unknown. We are again reminded of our quote of George Box: all models are wrong, but some are more useful than others [8]. This non-kinetic, steady state model of metabolism is wrong too, but it may be useful in predicting outcomes of alternate media compositions and cell line differences.

Making dynamics more accessible

The dynamics of biological systems and their components featured in some way in all chapters of this thesis. GEMMER (Ch. 4) enables users to cluster interaction networks based on the cell cycle phase where each component reaches its

expression peak. This facilitates understanding of how interaction networks may change due to the activity and presence of specific components over time.

In Ch. 3 we analyzed ChIP-exo experiments of Fkh1 and Fkh2 in two growth phases: the exponential phase and the stationary phase where glucose has run out. This enabled us to investigate differences in predicted target genes in these situations. Indeed there were substantial differences in the Fkh1 and Fkh2 targets between these two experiments. Furthermore, using the same data as utilized in GEMMER, we categorized the target genes both in terms of the cell cycle phase in which the expression was maximal ('peaked') and based on whether they displayed cyclic gene expression across cell cycles (i.e. whether they are cell cycle regulated or not). Finally, we investigated the temporal correlation of target gene expression with Fkh1 and Fkh2 expression across nine different datasets.

Ch. 5-7 dealt with predictions of system dynamics at steady state only, and then only in terms of an attracting steady state with constant behavior in time. Varying dynamics across time was discussed in Ch. 2 where we predicted system dynamics for 11 different network designs of the core cell cycle regulatory network in budding yeast and looked for oscillatory output. Although limit cycle oscillations are stationary (but not steady) state properties of a dynamical system, the resulting time course shows repeated oscillating dynamics of the components in the system. As we discuss in detail in Ch. 2, the alternatives to limit cycle oscillations include: (i) stable steady states (i.e. constant behavior across time as in Ch. 5-7), (ii) blowup (i.e. instability and concentrations increasing to infinity), (iii) checkpoint mechanisms or sequences of steady states attained by varying parameters and/or concentrations at specific moments during the simulation, or (iv) chaos. We can safely rule out (i), (ii) and (iv) since the cell cycle does not classify as a single steady state, nor as chaos, nor does it blow up: it is both oscillatory and reproducible. This leaves (iii) and limit cycles. It is unclear at the moment which of these two descriptions is most appropriate for the eukaryotic cell cycle. It is possible no final answer to this question exists, although in Ch. 2 we argue for the usage of limit cycles, first of all because it is most appropriate for the models discussed there and secondly because it does not rely on entrainment of the system by external oscillations such as the day/night cycle or mass growth and division.

Objectives

Our work in Ch. 2 discusses cell cycle oscillations in terms of limit cycles which are a stationary (but not steady) state property of the system. In this way, we model the cell cycle of yeast as having some inherent "objective" of being a continuously stable oscillation system. As we discussed in Ch. 2 this is an alternative way of looking at oscillation systems as opposed to viewing them as either a sequence of attracting steady states or a series of checkpoints to move through.

Ch. 3 did not explicitly deal with the topic of objectives. However, we can view our analyses there as silently assuming the "objective" of DNA-binding proteins being: to affect the expression of target genes. The underlying argument is Darwinian: the existence of a part of a protein that makes it bind to DNA is costly

in terms of resources that are used to make the amino acids and to stitch them together as a protein. Should the binding to the DNA not have a positive functional effect, the cost would result in selection against such a protein domain and its disappearance from the successful species. However, as discussed in Ch. 3 this view may be too simplistic. DNA binding does not necessarily result in altered expression patterns for nearby genes. However, binding is necessary for causing such an effect. Therefore, our analysis should only be taken as suggesting possible target genes, which still need to be verified experimentally using over-expression and deletion studies. Although not featured in this thesis, we have recently endeavored to do just this by using available over-expression [9] and deletion [10] experiments in addition to newly published ChIP data [11, 12] in order to expand on Ch. 3 and provide an even clearer picture of Fkh1 and Fkh2 functionality (Barberis & Mondeel, in press).

Ch. 5-7 deal with biological objectives by assuming that bacteria, such as *Cl. ljungdahlii*, and liver cancer cells (HCC) are optimized for biomass production, i.e. growth. This is a highly questionable assumption but allows us to cut through the sea of possible steady-state flux patterns that are possible on large metabolic networks and pick out the one that is optimal in some sense. For *Cl. ljunghdalii* and HCC cells we assumed the objective was growth through biomass production or maximizing ATP production. As unicellular organisms often compete by outgrowing each other, this may be reasonable. Instead for the glutathione detoxification subsystem we assumed the objective was to metabolize as much of paracetamol as possible. From the point of view that the same glutathione network should remove many xenobiotics from the organ that stands between the portal vein and the central blood circulation, this assumption may also be warranted. In all this demonstrates how biology may consider teleological causation.

8.2 Gear-shifting as a unifying lens

The concept of gear-shifting has emerged from the work in this thesis as a kind of unifying principle, or as a lens of sorts through which one can view many different biological phenomena.

As we proposed in Ch. 2 and 5, we envision that populations of cells consist of subpopulations possibly expressing differing phenotypes. Furthermore, point-mutations and shifts in gene expression provide mechanisms by which the functioning of any network interaction could be altered. Individual cells may thus be able to dynamically shift their network configurations resulting in an altered phenotype. Therewith the population should have a a phenotypic ''cloud'' rather than a well-defined phenotype. Such diversity in phenotype should allow a higher robustness and adaptability. Bacterial bet-hedging, early development and differentiation, drug resistance in cancer, and idiosyncrasy in drug effects, are but four examples where this may play a decisive role.

Our results in Ch. 2 indicate that, if these changes occur within the core cell cycle regulatory network, this should impact the ability of the network to exhibit oscillations. Therefore, differences in the affinity of Clb/Cdk1 complexes to bind

and phosphorylate Fkh transcription factors and, vice versa, in the affinity of Fkh to the CLB promoters may, paradoxically serve the function of enhancing robustness: diversity enhances robustness in many contexts [7]. Similarly, our results in Ch. 5 indicate that gear-shifting cofactors in the Wood-Ljungdahl pathway may result in different ATP and BHB yields coupled to acetogenesis.

Given that in slightly different circumstances other gear settings may be optimal, the concept of gear-shifting may explain the presence of multiple different versions of enzymes in the Wood-Lunghdahl pathway and of several cyclin molecules within the eukaryotic cell cycle. The various differing affinities for substrates cause different outcomes for these different gear settings and may allow them to serve as alternating sequential steps in a single process (the CDK/cyclin oscillations) or provide alternative yield ratios the optimality of which may change given varying circumstances such as the strength of thermodynamic backpressure (Ch. 5). Finally, by not altering the gear-settings too much they also serve to increase robustness since they may form backups in case a gene inhibition or deletion occurs that mutes the "optimal" gear setting. This is the better known explanation of the fitness benefit given by redundancy.

8.3 The next wave: data, scalable tools, and integrated models of multi-scale biology

The work in Ch. 2-8 of this thesis leaves many challenges and questions unanswered and suggests several exciting paths forward. From a high-level perspective there are clear avenues towards experimental validation of predictions and extension of existing datasets (Ch. 2, 3, 5 and 7), further refinement, application and testing of computational techniques (Ch. 2, 5, 6 and 7) and tools (Ch. 4) discussed in this thesis. Looking forward, all these paths forward will have to grapple with the multi-scale nature of biology and the large amount of information that is inherent to any living organism and now reaches us from the wide variety of sources, formats and techniques that are and will become available.

Specifically, Ch. 2 suggests experiments to investigate the existence of transcriptional inhibitory interactions of Sic1 with CLB2 and CLB3. As described in Ch. 2, such interactions were suggested by us in an attempt to describe the experimental evidence that Sic1 oscillations rescue viability of cells with low levels of mitotic Clb cyclins [13]: the model design incorporating a form of such interactions showed increased oscillatory potential compared to other designs. Ongoing efforts focus upon the hypothesis that this inhibitory interaction may occur through the Forkhead transcription factors of CLB2 and CLB3. Direct involvement of Sic1 as transcription repressor should constitute an exciting discovery since direct regulation of a gene promoter by a cyclin/Cdk inhibitor is unknown in budding yeast and should substantiate a mechanism coupling the end of the cell cycle to its start.

Ch. 3 calls for experiments validating the proposed roles of the genes that showed the strongest Forkhead binding signal, were the most reproducible across the three peak detection methods and/or hold the highest promise of interesting

biological insight. With a view toward elucidating the connection between the cell cycle and metabolism it may be prudent to focus on those target genes that have an important function in metabolism with both a high signal in ChIP-exo and a high reproducibility with respect to earlier experiments. Evidence of a significant effect of Fkh1 and/or Fkh2 on the transcription of such genes should position Fkh1 and/or Fkh2 in the connective "tissue" between the cell cycle and metabolism and should shed further light on how this connection influences the metabolic behavior of yeast.

Finally, through Tables 5.1 and 5.2, Ch. 5 suggests engineering targets in *Cl. ljungdahlii* for improving ATP yield coupled acetogenesis and for beta-hydroxybutyrate production, an interesting precursor for biodegradable plastics [14, 15].

On the tool development front, Ch. 2 indicates an opportunity to develop new or improved tools, e.g. extensions of the System Design Space toolbox [16] that allow user-friendly scanning of the design space in terms of dynamic properties, such as limit cycles or temporally defined regulation patterns. A much needed feature with regard to future large-scale, detailed and multi-scale models, will be to enable such simulations to take advantage of parallel computing on the user's machine or even in the "cloud" on servers. The SDS methodology is particularly suited to characterize properties of dynamical systems across vast areas of the parameter space. As such, it would be worthwhile to develop software to smoothly couple identified points of interest to other tools. For example, once a limit cycle point in the parameter space has been found using the SDS method, providing a smooth path to transfer the analysis to algorithms that can calculate its robustness region [17] should be useful in research.

Our work in Ch. 3 has highlighted the need for coming to terms with highly variable predictions in peak-detection methods. Until such differences in predictions are overcome, it may be best to analyze new datasets with multiple tools to get a complete picture of the signal in the data. It should be fruitful to investigate, using one or more verified sets of target genes, which (combination of) tools achieve the highest accuracy in predicting targets.

In Ch. 7 we elaborated a new method to generate flux balance analysis models based on the latest human metabolic map [18] that take transcriptomic data as well as medium/nutrient conditions into account. Our findings support the predictive potential of said basis since the predictions were, in most cases, matched by the experimental results. Our approach assumes that the $V_{\max}$ of a reaction is proportional to the level of the mRNA encoding the enzyme, with a proportionality constant equal for all reactions. This approach neglects differential translation and post-translational regulation, and assumes all k_{cat}'s to be equal. These assumptions can be improved upon in the future by consulting more literature information at first only for the steps that turn out to be important, and later perhaps by including information on the k_{cat} for each specific reaction.

Furthermore, with the continuing decreasing cost of omics experiments [19], it may become feasible to perform similar studies as in Ch. 7 whilst incorporating multiple transcriptomic datasets at multiple time steps. This should enable one to drop the highly simplified assumption of a constant transcriptome/expressome.

This, combined with dynamics of measured medium concentrations should significantly improve the realism of the dynamics underlying such models.

In terms of tool development, GEMMER (Ch. 4) leaves several avenues unexplored. GEMMER aims to answer, in a limited fashion, the ever-growing need of biologists to be able to look at the same data from multiple vantage points simultaneously. Pathways have an intuitive sequential structure, as illustrated by KEGG maps for example, but they are also collections of genes that form a network. These genes form interaction modules that have interactions with other such modules. In addition, gene products differ in abundance, temporal expression profile, transcription kinetics and biological function (kinases, transcription factors etc.). Zooming out, most genes express strongly in (only some) cell types, of (only a some) organisms, which interact in food-webs with other organisms. Ultimately, to truly deliver on the promise of (systems) biology for food technology, human health and animal welfare, climate change and ecological conservation we will need sophisticated ways to deal with information on all these levels.

It is unclear if organism-specific tools, like GEMMER, which are able to provide for niche functionality specific to one use-case, or organism agnostic tools, such as STRING [20] which allow easier switching and meta-genomic analyses, will prove a better way to go forward. Perhaps an integration between such tools is the way to go. However, all such tools should strive to move toward increasing the user's power to view the same data from multiple angles and to then integrate this with advanced querying capabilities, perhaps assisted by artificial intelligence. For any future iterations of GEMMER the goal should therefore be to add features that (more) seamlessly enable users to start at any single view: (a (KEGG) pathway, a database page, a hairball network, a model's interaction network based in SBML, or a depiction of the connections between biological modules such as the cell cycle and metabolism) and to transform the corresponding information to any of the other views. Tools could then suggest useful answers to complex queries such as: which genes in this biological module are most central (i.e. have the highest centrality measure scores)? Or, given known kinetic information, which enzymes have the highest control over the flux through a certain pathway? On the machine learning and artificial intelligence side, tools could be equipped with abilities such as suggesting (the latest) research results that are relevant to the further identification of the network and of aspects thereof that are being studied by the user, or even to suggest experiments. Tools with such features should become exquisitely useful in the hypothesis generation process and help in unlocking powerful and unintuitive truths that may currently be obfuscated by the data deluges.

There exists a more general challenge in computational biology, which has to do with coupling the massive amount of "big" data (e.g. from omics experiments as in Ch. 3) and large numbers of "big" models (Ch. 7) to the understanding we obtain from mechanistic bottom-up models [21] such as discussed in Ch. 2 and Ch. 6. Even though, budding yeast has many thousands of genes, each with many interactions (Ch. 3-4), some limited functionalities like cyclin/Cdk oscillations may be reproduced by relatively small computational systems. The crux lies both in the fact that beyond the boundaries of a small, bottom-up system lies the

rest of the system with some unspecified but nonetheless existing input-output relations, and in the missing connections and regulations that for the sake of the smaller model were assumed to be irrelevant or minimally relevant for the problem at hand. Current and future tools must therefore somehow find a balance associating the true genome-wide and meta-genomic nature of biology, which is complex, large, and perhaps not fully understandable by the human mind, and the need of the researcher and public to understand individual phenomena in terms of a limited number of "key" components. I especially foresee a role for machine learning and artificial intelligence in moving forward on this particular challenge since these methodologies could in theory perform functions like: (i) suggesting connections from a minimal model to the rest of the biological system, (ii) suggesting experiments to test certain model properties and predictions, (iii) distilling genome-wide networks into sub-networks for particular purposes and (iv) suggesting where small models are in contradiction with the current literature. For sure this will require enormous amounts of work in terms of annotating literature information better than it is done currently and coming up with constraints and design parameters for moving from small to large systems and back. But given the already enormous capabilities of machine learning in this regard, especially recommender systems that have popped up everywhere in our (technological) lives [22], these aims may well be achievable.

Ultimately, to utilize large multi-scale models of complete cells, organisms and communities, metabolic models such as those in Ch. 5 - 7 need to be coupled to models of other cellular processes such as the cell cycle, signaling, as well as to dynamics of the environment including other cell types and organisms within the same system or locale. Progress has been made in this regard by coupling multiple flux-balance analysis based cell-type models [23], coupling multi-algorithmic sub-models of a single bacterial cell [24] and by coupling a logical model of the cell cycle to a genome-wide steady state metabolic model of yeast [25]. However, both the computational challenge and experimental knowledge of the exact nature of connections between such processes leave much work to be done to achieve the goal of accurate, scalable and useful whole-cell modeling [26]. In order to live up to Erwin Schrödinger's statement that "The obvious inability of present-day physics and chemistry to account for [the events in space and time which take place within the spatial boundary of a living organism] is no reason at all for doubting that they can be accounted for by those sciences" [27], this whole-cell and whole-organism modeling endeavor should be a priority in terms of funding, time and energy investment in the decades to come.

8.4 The wisdom of the giants that came before

In closing, I'd like to come back to the quotes I inserted at the top of the chapters in this thesis and briefly look at the preceding chapters from their perspectives. I did not pick them in isolation and I feel that they form a coherent whole when viewed with the full content of this thesis.

As Douglas Adams wrote, the universe and life itself is inexplicable and one

should above all remember: "don't panic". Luckily, there is hope that by utilizing our ever increasing computational and mathematical abilities we can enable our knowledge of physics and chemistry to predict and perhaps understand the biochemical events happening inside organisms. Two critical aspects to understand are the cell cycle, which governs the process of cellular reproduction, and metabolism which governs the cellular chemical factory converting inputs into outputs and energy production and utilization. To achieve this we will need to make sense of vast amounts of data and extract from this a catalogue of all the parts of a biological cell, their connections and how these together produce the functionalities we live and observe. Part of the avenue to understanding biology and capitalizing on its potential for application to health, technology and ecological conservation, lies in building ever better models of sub-processes like Cdk/cyclin oscillations and in merging such models to form whole-cell models starting at microbes and moving up all the way to multi-cellular and multi-tissue models for humans. All such models will be inaccurate descriptions of the truth, but we should focus on continuously improving their agreement with experimental observation and their predictive powers and accuracy. This road may be never ending, but the further we progress along it the more we will tip the veil of the inexplicable nature of our own biology, of lives of the organisms around us, and thereby of not the least significant part of this cosmos.

References

[1] H. Kacser. "On Parts and Wholes in Metabolism". *The Organization of Cell Metabolism*. Ed. by G. R. Welch and J. S. Clegg. Boston, MA: Springer US, 1986, pp. 327–337. 10.1007/978-1-4684-5311-9_28.

[2] H. V. Westerhoff and D. B. Kell. "The methodologies of systems biology". *Systems Biology*. Elsevier, 2007, pp. 23–70. 10.1016/B978-044452085-2/50004-8.

[3] J. D. Davis, C. M. Kumbale, Q. Zhang, and E. O. Voit. "Dynamical systems approaches to personalized medicine". *Current Opinion in Biotechnology* 58 (2019), pp. 168–174. 10.1016/j.copbio.2019.03.005.

[4] L. Alberghina and H. V. Westerhoff. *Systems biology: definitions and perspectives*. Vol. 13. Springer Science & Business Media, 2007.

[5] M. Barberis *et al.* "Sic1 plays a role in timing and oscillatory behaviour of B-type cyclins". *Biotechnology Advances* 30 (2012), pp. 108–130. 10.1016/j.biotechadv.2011.09.004.

[6] L. Cai and B. P. Tu. "Driving the Cell Cycle Through Metabolism". *Annual Review of Cell and Developmental Biology* 28 (2012), pp. 59–87. 10.1146/annurev-cellbio-092910-154010.

[7] H. Kitano. "Biological robustness". *Nature Reviews Genetics* 5 (2004), pp. 826–837. 10.1038/nrg1471.

[8] G. E. P. Box and N. R. Draper. *Empirical model-building and response surfaces.* New York: John Wiley & Sons, 1987.

[9] S. R. Hackett *et al.* "Learning causal networks using inducible transcription factors and transcriptome-wide time series". *Molecular Systems Biology* 16 (2020), pp. 1–15. 10.15252/msb.20199174.

[10] P. Kemmeren *et al.* "Large-Scale Genetic Perturbations Reveal Regulatory Networks and an Abundance of Gene-Specific Repressors". *Cell* 157 (2014), pp. 740–752. 10.1016/j.cell.2014.02.054.

[11] M. J. Rossi *et al.* "A high-resolution protein architecture of the budding yeast genome". *Nature* 592 (2021), pp. 309–314. 10.1038/s41586-021-03314-8.

[12] O. Lupo, G. Krieger, F. Jonas, and N. Barkai. "Accumulation of cis - and trans - regulatory variations is associated with phenotypic divergence of a complex trait between yeast species". *G3 Genes | Genomes | Genetics* 11 (2021). Ed. by B. Andrews. 10.1093/g3journal/jkab016.

[13] S. J. Rahi *et al.* "The CDK-APC/C Oscillator Predominantly Entrains Periodic Cell-Cycle Transcription". *Cell* 165 (2016), pp. 475–487. 10.1016/j.cell.2016.02.060.

[14] M. Tortajada. "New waves underneath the purple strain". *Microbial Biotechnology* 10 (2017), pp. 1297–1299. 10.1111/1751-7915.12409.

[15] N. J. Claassens *et al.* "Harnessing the power of microbial autotrophy". *Nature Reviews Microbiology* 14 (2016), pp. 692–706. 10.1038/nrmicro.2016.130.

[16] J. G. Lomnitz and M. A. Savageau. "Design Space Toolbox V2: Automated Software Enabling a Novel Phenotype-Centric Modeling Strategy for Natural and Synthetic Biological Systems". *Frontiers in Genetics* 7 (2016), p. 118. 10.3389/fgene.2016.00118.

[17] M. Apri, J. Molenaar, M. de Gee, and G. van Voorn. "Efficient Estimation of the Robustness Region of Biological Models with Oscillatory Behavior". *PLoS ONE* 5 (2010). Ed. by D. Di Bernardo, e9865. 10.1371/journal.pone.0009865.

[18] E. Brunk *et al.* "Recon3D enables a three-dimensional view of gene variation in human metabolism". *Nature Biotechnology* 36 (2018), pp. 272–281. 10.1038/nbt.4072.

[19] R. Lowe *et al.* "Transcriptomics technologies". *PLOS Computational Biology* 13 (2017), e1005457. 10.1371/journal.pcbi.1005457.

[20] D. Szklarczyk *et al.* "The STRING database in 2011: functional interaction networks of proteins, globally integrated and scored". *Nucleic Acids Research* 39 (2011), pp. D561–D568. 10.1093/nar/gkq973.

[21] F. J. Bruggeman and H. V. Westerhoff. "The nature of systems biology". *Trends in Microbiology* 15 (2007), pp. 45–50. 10.1016/j.tim.2006.11.003.

[22] Z. Batmaz, A. Yurekli, A. Bilge, and C. Kaleli. "A review on deep learning for recommender systems: challenges and remedies". *Artificial Intelligence Review* 52 (2019), pp. 1–37. 10.1007/s10462-018-9654-y.

[23] I. Thiele *et al.* "When metabolism meets physiology: Harvey and Harvetta". *bioRxiv* (2018), p. 255885. 10.1101/255885.

[24] J. R. Karr *et al.* "A Whole-Cell Computational Model Predicts Phenotype from Genotype". *Cell* 150 (2012), pp. 389–401. 10.1016/j.cell.2012.05.044.

[25] L. van der Zee and M. Barberis. "Advanced Modeling of Cellular Proliferation: Toward a Multi-scale Framework Coupling Cell Cycle to Metabolism by Integrating Logical and Constraint-Based Models". *Yeast Systems Biology: Methods and Protocols*. Ed. by S. G. Oliver and J. I. Castrillo. New York, NY: Springer New York, 2019, pp. 365–385. 10.1007/978-1-4939-9736-7_21.

[26] A. P. Goldberg *et al.* "Emerging whole-cell modeling principles and methods". *Current Opinion in Biotechnology* 51 (2018), pp. 97–102. 10.1016/j.copbio.2017.12.013.

[27] E. Schrödinger. *What is life?* University Press: Cambridge, 1944.

Summary

With the dawn of "big data" in biology due to the advent of genome sequencing, biological and medical research has increasingly been reliant upon advanced data storage, data analysis and modeling techniques. Both the wide variety of organisms this planet supports and the complexity of the human body and the diseases it manifests, provide the rationale and highlight the need for systems approaches to biology. This need can be understood from the fact that emergent behavior can already arise in the nonlinear interaction of a very few components. The dependence of every organism or network component on most other organisms and/or network components of the same ecosystem or multicellular organism, and the vast expanse of those networks, provide the reasons for the study of that emergence to be essential for the science of biology. In recent decades, systems biology and bioinformatics, by integrating a systems perspective and mathematical modeling with experimental work and utilizing mathematics to sort though and organize the plethora of data produced by the latest omic technologes, respectively, have proved to be powerful at addressing numerous scientific questions. The combination of the capability to generate new, extensive and comprehensive experimental data with detailed analysis through computational methods promises to handle and gain understanding of the complexity, mechanisms and behavior of biological life on earth. In fact, due to that complexity, systems biology is arguably our only way forward.

This thesis discusses six scientific works within the fields of systems biology and bioinformatics, covering a wide spectrum of biological questions and organisms. The overarching theme espoused here is that by bringing a handful of computational methodologies in contact, at times after extending them, with new and quantitative experimental data, new principles can be proposed and/or tested that would not otherwise have been discovered. Chapters 2-7 are unified in their conception of the cell as a system of integrated fluxes of mass and information, in the application of computational approaches to answer the questions at hand and in their aim for computation to drive new (experimental) biological discoveries. In the course of Ch. 2-7 we present new datasets, provide novel tools, develop new models, propose novel (extensions of) computational methodologies and rationalize and assess existing such methodologies. As such, this thesis provides a glance into the cutting-edge of biomedical research in this data-driven, computation-assisted age.

Arguably, there is no more fundamental process in life than life giving rise to more of itself by progression through the cell cycle and ultimately to the birth of new cells in cell division. The first three chapters (Ch. 2-4) arose from an interest in improving our understanding of the progression of the budding yeast cell cycle as a continuous process with neither end nor beginning, and how it interfaces with other cellular processes such as metabolism and signal transduction. Specif-

ically, our interest goes out to the waves-of-cyclins phenomenon at the heart of the cell cycle regulatory network conserved throughout eukaryotes. We investigate both the influence of network properties on this oscillatory phenomenon (Ch. 2) and how this process may coordinate with or drive other cellular processes through two specific transcription factors (Ch. 3). There exists a general need to understand and visualize the multitude of interactions between cellular components, identified using a wide variety of experimental and computational approaches, across functional, temporal and concentration scales. This calls for appropriate visualization tools and databases and this topic forms the subject of Ch. 4.

Specifically, in Ch. 2 we investigate a family of network designs and kinetic models of the budding yeast cell cycle regulatory network revolving around three cyclin/Cdk1 complexes and their common stoichiometric inhibitor Sic1. We show how network design modifications influence the ability to yield sustained oscillations in the form of limit cycles. To do so, we implemented and extended a method recently proposed by Michael Savageau and colleagues to partition the parameter space of any kinetic model into distinct regions with unique characteristics of that model's behavior. We applied a sampling approach to this *Design space* and as a consequence we obtained many different parameter sets that yield oscillatory behavior across vastly different areas of the parameter space and across 7 different network designs. Our results suggest that specific cyclin(Clb3)-centered regulation is pivotal for obtaining stable oscillations. We further reinforce the existing understanding of the importance of positive and negative feedback loops between the oscillating components and suggest specific subsets of these that seem especially relevant. Several experimentally testable hypotheses emerge from this work.

In Ch. 3, we zoom in on the forkhead transcription factors Fkh1 and Fkh2, which play important roles in the cell cycle regulatory network and especially in the waves-of-cyclins phenomena investigated in Ch. 2. Specifically, we present a novel dataset from ChIP-exo experiments which allows to retrieve the specific binding sites at gene promoters where Fkh1 and Fkh2 anchor. We analyze this new dataset in the context of multiple previous studies into these transcription factors. We make use of three different data analysis methods for peak detection, including the newly proposed *maxPeak* method. Using the different peak-detection methods we obtained highly different binding events. This result suggests that simultaneously utilizing multiple peak-detection approaches enhances the assessment of the (limited) accuracy of retrieved binding events. Our work, as a result of using all three methods simultaneously, suggests that the forkhead transcription factors bind genomic locations upstream of many cell cycle-related genes as well as of genes involved in metabolism and/or signal transduction. We propose based on these findings that the Forkheads may be hubs integrating the cell cycle and metabolism.

In Ch. 4, we develop a new web-based database and visualization tool (GEMMER) which integrates a variety of sources of information concerning all protein-coding genes in budding yeast and allows users to craft specific and informative visualizations of the topology of interaction networks. The database behind

GEMMER has been used in analyzing the ChIP-exo data from Ch. 3.

The last three chapters of this thesis focus on that other process by which life produces more of itself: metabolism. Specifically, our focus is on metabolic networks in acetogenic bacteria (Ch. 5) and human liver (Ch. 6-7).

Specifically, Ch. 5 is concerned with the concept of gear-shifting: an organism's hypothetical ability to express metabolic enzymes that result in different stoichiometric yields in order to navigate a trade-off between rate and yield. As an example we will discuss acetogenesis: Producing more ATP at the same ATP/ADP ratio, coupled to acetogenesis, reduces Gibbs free energy dissipation and as a consequence decreases reaction rates. Thereby acetogenic organisms face a rate/efficiency trade-off. In particular we extend the non-equilibrium thermodynamics of coupled processes to arrive at a generalized *variomatic gear-shifting* principle. We then discuss how acetogenic bacteria may be able to shift gears depending on the environmental conditions they are presented with. As this gear shifting corresponds to a variation in the flux pattern through an existing metabolic network, we use flux balance analysis simulations to highlight that one such acetogen, *Clostridium ljungdahlii*, is at least in theory able to gear-shift in the acetogenesis process.

Ch. 5 deals with metabolism exclusively in terms of fluxes. However, concentrations are also important, e.g. to understand regulation. It has been a challenge to deal with concentrations without losing the advantage of flux balance analysis that it only requires network topology and not enzyme kinetic information. In Ch. 6 and 7 we discuss two approaches to deal partially with concentrations in FBA, i.e. in terms of serum concentrations of biomarkers (Ch. 6) and in terms of a relationship between uptake fluxes of medium metabolites and their concentrations (Ch. 7).

Ch. 6 discusses ways to use serum biomarkers to predict the capacity of glutathione-mediated detoxification in human liver utilizing (i) non-dynamic genome-scale metabolic models through flux variability analysis or (ii) a dynamic kinetic model. This detoxification is relevant for the levels of Reactive Oxygen Species (ROS) as well as of xenobiotics, including medicinal drugs. The aim of Ch. 6 was to find out if a previously proposed biomarker prediction algorithm could match predictions obtained from a previously published kinetic model. In the process, we computationally re-implement and validate (for a rather limited set of network structures) the original technique. We also rationalize the methodology by providing a proof of its validity for a limited network topology. Furthermore, we correct the scientific record on some predictions made by the kinetic model. Ultimately, our analysis shows that for the specific glutathione conjugation network studied here, the biomarker prediction algorithm is not capable of matching the predictions from the kinetic model. This proves that FBA-biomarker prediction methodology may well fall short for this and other biomedically important systems.

In Ch. 7 we investigate whether two hepatocellular carcinoma cell lines behave identically, or whether the cancer cells may differ in gear setting, or in gear shifting capability. Cancer cells tend to have a low gear setting in terms of ATP produced per glucose molecule, by fermenting to lactate rather than oxidizing to

carbon dioxide, and may even switch gears between glucose and glutamine as substrate. These two phenomena are known as the Warburg and WarburQ effect, respectively. The experimental results show that the hepatocellular carcinoma cell lines are neither addicted nor shifting to glutamine (i.e. show no WarburQ effect) and show only a moderate Warburg effect. Significant aerobic glycolysis to lactate occurs but it is not as much as in a complete gear shift to lactate. These observations suggest a potentially large respiratory flux. Furthermore, the two cell lines show a difference in growth rates across media, responding especially to glucose, although both show an insensitivity to glucose concentrations in the medium when they exceed a certain level. Also, we observe that growth rates for both cell lines are independent of glutamine in media containing glucose (again a lack of the WarburQ effect) and growth rates on glutamine alone are much lower than those on glucose alone. We complement this experimental investigation by developing a new flavor of flux balance analysis that integrates metabolite concentrations in the medium and transcriptomics data from the cell lines. Through this novel method, cell nutrition and gene expression can now be dealt with simultaneously by FBA through a scaling factor. First, we highlight based on existing transcriptomics data that significant differences exist between the metabolic enzymes of the two cell lines and their resulting Reaction Activity Scores (RAS). Second, we use the "Ansatz" that to some extent variations in medium metabolite concentrations can be modelled in FBA by setting uptake bounds proportional to the concentrations of the corresponding metabolites. Thirdly, we introduce a parameter through which we can adjust the relative control on growth rate exercised by medium metabolite concentrations and transcriptome. The value of that parameter is then "fitted" to the experimental data and complete flux distributions are predicted that were not accessible experimentally. The computational results of growth capabilities and glucose insensitivity above a threshold track those of the experimental characterization and suggest that respiratory flux is important in these cell lines under low or zero glucose conditions: gear shifting to lactate production is incomplete, suggesting the presence of smooth variomatic gear-shifting as opposed to shifting between two states (exclusive lactate production and no lactate production).

Finally, in Ch. 8 we discuss the work in Ch. 2-7 in terms of four unifying concepts, i.e. network wiring, adjustable coupling, dynamics and biological objectives, in terms of the big data explosion in society and in terms of the future of genome-wide (big data), mechanistic models in biology. We use the four concepts to hypothesize on possibilities to extend the data, tools and methods proposed in Ch. 2-7. The next wave of research will need to focus on fluid switching between multiple viewpoints on the same data, and to integrate machine learning and artificial intelligence in the hypothesis generation and testing processes themselves, and come to terms with the general aspects of "gear-shifting" present throughout molecular biology. It may thereby begin to fulfill the promises that (systems) biology holds for biotechnology, healthcare and ecology.

Samenvatting

Met de komst van "big data" in de biologie ten gevolge van de komst van DNA-nucleotidevolgordebepaling voor het gehele genoom, is biologisch en medisch onderzoek in toenemende mate afhankelijk geworden van geavanceerde gegevensopslag, gegevensanalyse en modelleringstechnieken. Tegelijkertijd veroorzaken zowel de grote verscheidenheid aan organismen op aarde als de complexiteit van het menselijk lichaam en de ziekten die het vertoont, een behoefte aan systeem- of netwerkbenaderingen van de biologie. Deze behoefte kan worden begrepen uit het feit dat emergentie al kan ontstaan uit de niet-lineaire interactie van een zeer klein aantal componenten en dus zeker indien dat aantal in de duizenden loopt. De onderlinge afhankelijkheid van elk organisme of netwerkcomponent van (vrijwel) alle andere organismen en/of netwerkcomponenten vormt de basis voor dit fenomeen in de biologie. In de afgelopen decennia zijn systeembiologie en bio-informatica, respectievelijk door het systeemperspectief en de wiskundige modellering te integreren met experimenteel werk of door gebruik te maken van de nieuwste polynucleotidensequencing technologieën, krachtig gebleken bij het onderzoeken van tal van wetenschappelijke vragen. De combinatie van het vermogen om grote hoeveelheden nieuwe gegevens te genereren met het uitvoeren van gedetailleerde analyses door middel van computationele methoden schept de mogelijkheid om complexe mechanismen en het gedrag van biologisch leven op aarde te onderzoeken en te begrijpen, en is wellicht onze enige manier om dit te bereiken.

Dit proefschrift bespreekt zes wetenschappelijke werken op het gebied van systeembiologie en bio-informatica, die een breed spectrum van biologische vraagstukken en organismen bestrijken. Het overkoepelende thema is dat met behulp van een select aantal computationele methodologieën, vaak gecombineerd met nieuwe en kwantitatieve experimentele gegevens, nieuwe principes kunnen worden geïdentificeerd en/of getest. Hoofdstukken 2-7 zijn verenigd in hun conceptie van de cel als een systeem van geïntegreerde fluxen van massa en informatie, in de toepassing van berekeningen om de hoofdvragen te beantwoorden en in hun streven naar nieuwe (experimentele) biologische ontdekkingen. In de loop van hoofdstuk 2-7 presenteren we nieuwe gegevensbestanden, voorzien we in nieuwe methodes en in nieuwe modellen, stellen we nieuwe (uitbreidingen van) computationele methodologieën voor en beoordelen we bestaande rekenmethodes. Als zodanig biedt dit proefschrift een blik op het moderne biomedische onderzoek in dit gegevensgestuurde, computerondersteunde tijdperk.

Er is wellicht geen fundamenteler proces binnen de biologie dan het proces waarin levende dingen meer van zichzelf voortbrengen door middel van de celcyclus en celdeling. De eerste drie hoofdstukken (hoofdstuk 2-4) kwamen voort uit een interesse in de celcyclus van *Saccharomyces cerevisiae* (bakkersgist) als een continu proces, alsmede in hoe deze cyclus samenwerkt met andere cellulaire

processen zoals metabolisme en signaal-transductie. In het bijzonder gaat onze interesse uit naar het fenomeen van golving van cyclines dat centraal staat in het regulerende netwerk van de celcyclus en geconserveerd is in eukaryoten. We onderzoeken zowel de invloed van netwerkeigenschappen op de mogelijkheid om dit oscillerende fenomeen te vertonen (hoofdstuk 2) als hoe de celcyclus andere cellulaire processen kan coördineren, en op zijn beurt door die processen aangestuurd kan worden via twee specifieke transcriptiefactoren (hoofdstuk 3). De brede behoefte om de veelheid aan interacties tussen cellulaire componenten die zich afspelen op verschillende functionele, temporele en concentratieschalen te begrijpen en te visualiseren, resulteert in de behoefte aan geschikte visualisatiemethodes en gegevensbestanden. Deze behoeften vormen het onderwerp van hoofdstuk 4.

In hoofdstuk 2 onderzoeken we een familie van kinetische modellen rond drie cycline/Cdk1-complexen en hun gemeenschappelijke stoichiometrische remmer Sic1. We laten zien hoe wijzigingen in het netwerkontwerp van invloed zijn op het vermogen om aanhoudende oscillaties op te leveren in de vorm van limietcycli. Om dit te doen, hebben we een recent door Savageau en collegae voorgestelde methode geïmplementeerd en uitgebreid om de parameterruimte van de kinetische modellen onder te verdelen in verschillende gebieden met unieke kenmerken. Steekproefsgewijs hebben we verscheidene parametercombinaties verkregen die tot zulke continue oscillaties leidden. Deze parametercombinaties bevonden zich in ver van elkaar gelegen delen van de parameterruimte. Dit gold voor zeven verschillende netwerkstructuren. Deze resultaten suggereren dat slechts een beperkt aantal types cycline (Clb3) regulatie stabiele oscillaties genereren. We versterken verder het bestaande begrip van het belang van positieve en negatieve terugkoppelingen tussen de oscillerende componenten en suggereren dat sommige daarvan van bijzonder belang zijn. Dit werk brengt ons tot het formuleren van verschillende experimenteel toetsbare hypothesen.

In hoofdstuk 3 richten we ons op de forkhead-transcriptiefactoren Fkh1 en Fkh2, die een belangrijke rol spelen in het netwerk dat de celcyclus reguleert en vooral op het verschijnsel van cyclinegolven zoals die onderzocht werden in hoofdstuk 2. Specifiek presenteren we een nieuwe serie gegevens afkomstig uit ChIP-exonuclease-experimenten die het mogelijk maakt om te vinden waar Fkh1 en Fkh2 aan de genpromotoren binden. We analyseren deze nieuwe gegevens in samenhang met eerdere studies naar deze transcriptiefactoren en maken gebruik van drie verschillende gegevensanalysemethoden voor piekdetectie. Analyse met behulp van deze verschillende piekdetectiemethoden bracht grote verschillen aan het licht tussen de door deze methodes gesuggereerde bindingsplaatsen. Dit resultaat suggereert dat gelijktijdig gebruik van meerdere benaderingen voor piekdetectie gunstig is om (het gebrek aan) de nauwkeurigheid van de gevonden bindingsplaatsen vast te stellen. Ons werk suggereert dat de forkhead-transcriptiefactoren stroomopwaarts van zowel veel celcyclusgerelateerde genen als van genen die betrokken zijn bij stofwisseling of signalering binden aan het DNA. Op basis van deze bevindingen stellen we voor dat de *Forkheads* knooppunten zijn waarlangs celcyclus- en stofwisselingsregulatie samenkomen.

In hoofdstuk 4 introduceren we een nieuw webgebaseerd gegevensbe-

stand en visualisatiehulpmiddel (GEMMER) dat een verzameling aan informatiebronnen over alle eiwitcoderende genen van *Saccharomyces cerevisiae* combineert en waarmee gebruikers specifieke en informatieve visualisaties van de topologie van interactienetwerken kunnen maken. Het gegevensbestand achter GEMMER is tevens gebruikt bij het analyseren van de ChIP-exo-gegevens in hoofdstuk 3.

De laatste drie hoofdstukken van dit proefschrift richten zich op dat andere proces waardoor het leven meer van zichzelf produceert: de stofwisseling (ook wel metabolisme). Onze focus ligt in het bijzonder op stofwisselingsnetwerken in acetogene bacteriën (hoofdstuk 5) en menselijke lever (hoofdstuk 6-7).

Hoofdstuk 5 bespreekt het concept van "schakelen": het vermogen van een organisme om zich te voorzien van stofwisselingsenzymen die resulteren in verschillende stoichiometrische opbrengsten, om daardoor een afweging te maken tussen snelheid en opbrengst van ATP-synthese. Als voorbeeld bespreken we azijnzuurvorming (acetogenese). Het produceren van meer ATP bij dezelfde ATP/ADP-verhouding, gekoppeld aan acetogenese, vermindert dissipatie van Gibbs vrije energie en als gevolg daarvan neemt de reactiesnelheid af. Daardoor hebben acetogene organismen te maken met een afweging tussen snelheid en efficiëntie. In het bijzonder breiden we de niet-evenwichtsthermodynamica van gekoppelde processen uit met een algemeen *variomatisch schakel*-principe. Vervolgens bespreken we hoe acetogene bacteriën kunnen schakelen, afhankelijk van de omgevingscondities waarmee ze worden geconfronteerd. Aangezien dit schakelen overeenkomt met een variatie in het stromingspatroon over een bestaand stofwisselingsnetwerk, gebruiken we simulaties met behulp van fluxbalansanalyse (FBA) om te benadrukken dat een dergelijk acetogeen micro-organisme, *Clostridium ljungdahlii*, in ieder geval in theorie in staat is om te schakelen in het acetogenese proces.

Hoofdstuk 5 gaat uitsluitend over metabolisme in termen van stromen. Concentraties zijn echter ook belangrijk, o.a. om regulatie te begrijpen. Het is een uitdaging om met concentraties om te gaan binnen het kader van FBA waar het alleen om netwerktopologie gaat en niet over kinetische informatie betreffende enzymen. In hoofdstuk 6 en 7 bespreken we twee benaderingen om gedeeltelijk om te gaan met concentraties in FBA, te weten in termen van serumconcentraties van bioindicatoren (hoofdstuk 6) en in termen van een relatie tussen opnamestromen van voedingsstoffen uit het medium en hun concentraties (hoofdstuk 7).

Hoofdstuk 6 bespreekt manieren om serumbioindicatoren te gebruiken om de capaciteit van glutathion-gemedieerde ontgifting in de menselijke lever te voorspellen met behulp van (i) niet-dynamische metabole modellen op genoomschaal door middel van fluxvariabiliteitsanalyse en (ii) een dynamisch kinetisch model. Deze ontgifting is relevant voor de niveaus van *Reactive Oxygen Species* (ROS) en van xenobiotica, waaronder geneesmiddelen. Het doel van hoofdstuk 6 was om uit te zoeken of een eerder voorgesteld voorspellingsalgoritme voor bioindicatoren dezelfde indicatoren zou opleveren als een eerder gepubliceerd kinetisch model. Hierbij implementeren en valideren we de oorspronkelijke rekentechniek opnieuw (voor een beperkt aantal netwerkstructuren). We rationaliseren ook de methodologie door een bewijs te leveren van de geldigheid

ervan voor een beperkte netwerktopologie. Verder corrigeren we enkele van de eerder door het kinetische model gedane voorspellingen. Uiteindelijk laat onze analyse zien dat specifiek voor het glutathionconjugatienetwerk dat hier is bestudeerd, de fluxbalansanalysemethode niet in staat is om de voorspellingen van het kinetische model te evenaren. Dit toont aan dat de FBA methodologie tekortschiet in het betrouwbaar voorspellen van bioindicatoren, zeker in het farmacologisch belangrijke glutathion systeem, maar waarschijnlijk ook in algemenere zin.

In hoofdstuk 7 onderzoeken we of twee levercarcinoomcellijnen zich identiek gedragen, of dat de kankercellen kunnen verschillen in "versnelling" of in "schakelvermogen". Kankercellen hebben de neiging om een lage versnelling te hebben in termen van ATP geproduceerd per glucosemolecuul, en kunnen zelfs schakelen tussen het volledig oxideren van glucose en hetzij het fermenteren van glucose tot melkzuur, hetzij het deels afbreken van glutamine to melkzuur. Deze twee verschijnselen staan bekend als respectievelijk het Warburg- en WarburQ-effect. De experimentele resultaten laten zien dat deze levercarcinoomcellijnen niet verslaafd zijn aan glutamine (d.w.z. geen WarburQ-effect vertonen) en slechts een matig Warburg-effect vertonen. Aanzienlijke aerobe glycolyse tot melkzuur vindt plaats, maar het is niet zodanig als bij een volledige omschakeling naar melkzuur als eindproduct. Deze waarnemingen suggereren een mogelijk grote respiratoire flux. Bovendien vertonen de twee cellijnen een verschil in groeisnelheid tussen media, waarbij ze vooral reageren op glucose. Beide vertonen een ongevoeligheid voor veranderingen in glucoseconcentraties in het medium als deze eenmaal een bepaald niveau ontstijgt. We zien ook dat de groeisnelheden voor beide cellijnen onafhankelijk zijn van glutamine in media die glucose bevatten (een gebrek aan het WarburQ-effect) en dat de groeisnelheden op glutamine alleen veel lager zijn dan die op glucose alleen.

We vullen dit experimentele onderzoek aan door een nieuwe smaak van fluxbalansanalyse te ontwikkelen die metabolietconcentraties uit het medium met transcriptoomgegevens van de cellijnen integreert. Door deze nieuwe methode kunnen celvoeding en genexpressie nu gelijktijdig door FBA worden behandeld door middel van een weegfactor. Ten eerste benadrukken we op basis van bestaande transcriptoomgegevens dat er significante verschillen bestaan tussen de stofwisselingsenzymen van de twee cellijnen, althans tussen de hoeveelheden ervan. Ten tweede gebruiken we een aanpak waarmee variaties in gemiddelde metabolietconcentraties tot op zekere hoogte in FBA kunnen worden gemodelleerd door opnamegrenzen in te stellen die evenredig zijn aan de concentraties van de overeenkomstige metabolieten. Ten derde introduceren we een parameter waarmee we de relatieve controle op de groeisnelheid die wordt uitgeoefend door metabolietconcentraties uit het medium en transcriptoom kunnen aanpassen. De waarde van die parameter wordt dan 'aangepast' aan de experimentele gegevens en er worden volledige fluxverdelingen voorspeld die experimenteel nog niet gemeten werden. De rekenresultaten van groeimogelijkheden en glucose-ongevoeligheid boven een bepaalde waarde zijn in overeenstemming met de experimentele karakterisering en suggereren dat respiratoire flux belangrijk is in deze cellijnen onder lage of nul glucose-omstandigheden: het

schakelen naar melkzuurproductie is onvolledig, wat wijst op de mogelijkheid tot soepel schakelen in tegenstelling tot het slechts kunnen schakelen tussen twee toestanden (exclusieve melkzuurproductie of geen melkzuurproductie).

Tot slot bespreken we in hoofdstuk 8 het werk uit hoofdstukken 2-7 in termen van vier verenigende concepten: netwerktopologie, instelbare koppeling tussen processen, dynamiek en biologische doelstellingen, alsook in het licht van de overvloed aan gegevens in de samenleving en de toekomst van genoombrede, mechanistische modellen in de biologie. We gebruiken deze concepten om uitbreidingen te suggereren aan de gegevens, hulpmiddelen en methoden die werden voorgesteld in hoofdstukken 2-7. Toekomstig onderzoek zal er goed aan doen om zich onder andere te concentreren op het feilloos schakelen tussen meerdere gezichtspunten op dezelfde gegevens alsook op het integreren van 'machine learning' en kunstmatige intelligentie. Het dient daarbij te gaan zowel om het genereren als het testen van hypothesen om zo de kracht van (systeem)biologie in biotechnologie, gezondheidszorg en ecologie volledig tot uiting te laten komen.

Acknowledgments

> "Bernard of Chartres used to compare us to [puny] dwarfs perched on the shoulders of giants. He pointed out that we see more and farther than our predecessors, not because we have keener vision or greater height, but because we are lifted up and borne aloft on their gigantic stature."
>
> — John of Salisbury [1]

Looking back on the past several years during which this thesis came about, I felt drawn toward the often quoted phrase originally attributed to Bernard of Chartres displayed above. Indeed, working to advance science or understanding in any field, however small the step forward, is to stand on the shoulders of giants that have come before. However, in my experience, one also stands in the presence of present-day giants, without whom none of it would have happened or mattered. During the past several years it has been my fortune and blessing to be surrounded by many people who have given me their time, assistance and companionship. Although no words can do justice to their importance, I attempted a first-order approximation below.

I would first and foremost like to express my sincere gratitude to my promotor, Hans Westerhoff. Hans, you enlightened me with all the scientific support and supervision that a graduate student can expect from their professor. To stand on the shoulders of giants, you must first accept that you don't know everything. Needless to say, I did not at the start of my PhD, nor do I now, know everything or even a lot, but perhaps a little. As Einstein once said: "as our circle of knowledge expands, so does the circumference of darkness surrounding it." To be sure, my circle of knowledge and my resulting circumference of darkness and ignorance would not have grown as much without your incessant ability to pinpoint the most painful flaws and interesting aspects of any scientific problem we were working on. You have influenced all the chapters in this work. Even on the published chapters for which you are not technically a co-author, you should be considered a *co-author in spirit*. I hope you continue to increase our collective circle of knowledge and our perimeter of ignorance for a long time to come.

Second, I would like to thank my other promotor, Matteo Barberis. Matteo, you are the person I spent the most time with during the completion of the work contained in this thesis, first in person and after your move to the UK on video calls, and they were good times indeed! I know I learned a lot from our time and work together and I think the results bore fruit. Some time-consuming projects we worked on together did not ultimately make it into this thesis, but I hope you are happy with the result nonetheless. Those too will make it out into the world someday I am sure. I wish you all the best on the road to come.

Third, looking back, I would like to acknowledge the teachers I spent the

most time with during my master's degree in mathematical biology: Bob Planqué, Frank Bruggeman and Joost Hulshof. It is due to your contagious enthusiasm that I was interested in doing research in the first place: I hope you keep inspiring more students to do the same.

As for my other colleagues, I would first like to acknowledge Stefania Astrologo whose birthday gift of Schrödinger's book inspired the opening pages of this thesis. Stefania, we spent (too) many hours tracking down Python and R bugs, trying to figure out confusing model output and being frustrated together. I learned a lot from those sessions though and I look back on them fondly. I hope you do too! Finally, please forgive my Dutch directness when it wasn't wanted in the past, and thank you for occasionally nagging at me to finally finish this thesis! It may never have been completed otherwise.

Christian Linke, you probably served as something approximating my psychologist during our cappuccino breaks at the start of my time at the UvA and I sincerely thank you for that. Unfortunately, you weren't there for its completion, but I hope life treats you well in Germany and I'm always available for another cappuccino if you find yourself back in Amsterdam or Lisse.

I would like to thank Ewelina Węglarz-Tomczak for the opportunity to collaborate on the HCC project featured in Ch. 7 during the end of my PhD contract. This project was a breath of fresh air that renewed my enthusiasm when I was nearing the end of my other projects. I hope the rest of your time in Amsterdam, together with Jakub, will be productive, fun and filled with an abundance of equestrianism!

To my other former colleagues (Will, Yanfei, Miguel, Jihed, Mihaly, Yanhua, Anchal, Rogier, Raju, Diewertje, Jessica, Ablikim, Rurika, Lucas, Till, Blaise, Jana, Mannus, José, Ilona, Isa and others): thank you for the fun times we had together playing pool, devouring hamburger and drinking beer at the Oerknal after work. And of course the, not infrequent, unproductive but enjoyable times we had at the office.

Over the years I was fortunate to have to pleasure of supervising several students in our lab, who at times taught me more than I taught them: Victor Ubels, Vivian Ogundipe, Willem van Dorp, Willem Kiel, Dennis Melkert, Ivar van Galen and Eva Mooij. I apologize for any shortcomings on my behalf in supervising you and thank you for all that you taught me. Some of your work features in this thesis and would not have been done without your assistance. I know some of you are eyeing or already pursuing further scientific careers and I'm proud to have perhaps made a (small) contribution to that end.

Perhaps most importantly, I would like to thank my wife Emma and family (Harry, Janny, Antoinella, Dominique, Isabella, Sofia, Wout, Iet, Daan, Diana, Liene and Fie) for their presence and support during these years. Dominique was instrumental to the existence of Ch. 4, but all of them were instrumental to the existence of a thesis in the first place. Zonder jullie was het niet gebeurd en had het ook niets uitgemaakt. I need to especially acknowledge my wife Emma in this regard, who was not yet my wife at the start of all of this. Emma, you stuck with me through late nights when I was still working on my laptop, proof-read many parts of this thesis and believed in me all the way. I will do my best to reciprocate

this love and support for the rest of our lives.

Finally, I would like to acknowledge, in no particular order or grouping and for multiple reasons, the following people for helping me substantially along the way to this thesis: Chiara Damiani, Marina Wright Muelas, Robert Todd, Jason Lomnitz, Wolfram Liebermeister, Samrina Rehman, Malkhey Verma, Frédéric Crémazy and Petter Holland.

To those giants that I failed to thank above: my deepest apologies and know that I appreciate you regardless.

Chapter-by-chapter acknowledgments

In closing, I list acknowledgments and contributions on a chapter-by-chapter basis, on behalf of myself and the co-authors of each chapter, for assistance with the content of each chapter and the funding thereof.

Ch. 2 - Waves

Acknowledgments. The authors thank Michael A. Savageau for helpful discussions and Jason Lomnitz for help with the System Design Space toolbox, Irene Zorzan for checking the mathematical derivation of the quasi-steady-state approximation, and the anonymous reviewers for insightful comments. This work was supported by the Systems Biology Grant of the University of Surrey to M.B., and the SILS Starting Grant of the University of Amsterdam to M.B.

Contributions. M.B. formulated the experimental rationale underlying the study, and conceptualized, conceived and designed the study. T.D.G.A.M. and M.B. designed the computational analysis, which was performed by T.D.G.A.M. O.I. implemented the first sets of autonomously oscillating cell cycle models. H.V.W. helped with the explicit check for the presence of complex conjugate eigenvalues. W.L. and M.B. implemented the quasi-steady-state approximation. M.B. provided the biological interpretation of network designs and of computational analyses. M.B. and T.D.G.A.M. analyzed the data. M.B. wrote the paper with contributions from T.D.G.A.M., W.L. and H.V.W. T.D.G.A.M. wrote the Methods section and Supplementary Information with contributions from M.B. M.B. provided scientific leadership and supervised the study.

Ch. 3 - ChIPs

Acknowledgments. The work in Ch. 3 was supported by the Systems Biology Grant of the University of Surrey to M.B.; the Swammerdam Institute for Life Science Starting Grant of the University of Amsterdam to M.B.; and the Novo Nordisk Foundation, Vetenskapsrådet, and Knut and Alice Wallenberg Foundation to J.N. The authors thank the anonymous reviewers for insightful comments.

Contributions. M.B. conceived and designed the study. P.H. and M.B. designed the experimental analysis, which was performed by P.H. P.H. implemented the peak detection pipeline with contributions from T.D.G.A.M. and M.B. M.B. and T.D.G.A.M. designed the data analysis, which was performed by T.D.G.A.M.

T.D.G.A.M. and M.B. analyzed the data. M.B. provided the biological interpretation of data. M.B. and T.D.G.A.M. wrote the paper, with contributions from P.H. and J.N. M.B. provided scientific leadership and supervised the study.

Ch. 4 - GEMMs

Acknowledgments. We thank Brenda J. Andrews and her lab members for providing the corrected Excel files linked to the CYCLoPs database; Dominique Groenveld for assistance with the server setup; Paul Verbruggen for help with the final layout of the GEMMER logo; and Lucas van der Zee for valuable discussions.

Contributions. M.B. conceived the idea, the tool name and designed the strategy of the study. M.B., T.D.G.A.M. and F.C. designed the tool and its features. T.D.G.A.M. and F.C. programmed the source code. T.D.G.A.M. and M.B. implemented the tool features. T.D.G.A.M. and M.B. wrote the paper, with contributions from F.C. M.B. provided scientific leadership and supervised the study.

Ch. 5 - Gears

Acknowledgments. The authors wish to acknowledge Olga Revelles, Stéphanie Follonier and Tanja Narancic for key discussions during the early stages of this work. The research leading to these results received funding from the European Union Seventh Framework Programme (FP7/2007-2013) under Grant Agreement No. 311815 (SYNPOL project).

This work was financially supported by the Netherlands Organization for Scientific Research (NWO) in the integrated program of WOTRO [W01.65.324.00/project 4] Science for Global Development as well as by various systems biology grants, including Synpol: EU-FP7 [KBBE.2012.3.4-02 #311815], Corbel: EU-H2020 [NFRADEV-4-2014-2015#654248], Epipredict: EU-H2020 MSCA-ITN-2014-ETN: Marie Skłodowska-Curie Innovative Training Networks (ITN-ETN) [#642691], and BBSRC China [BB/J020060/1].

Contributions. H.V.W. conceived the study and provided scientific leadership. T.D.G.A.M., A.A. and H.V.W. wrote the manuscript. Modeling and data analysis was performed by T.D.G.A.M., S.A., Y.Z. and H.V.W. I.G., M.B., P.D., M.V. and S.R. provided relevant input throughout the study.

Ch. 6 - Markers

Acknowledgments. The authors would like to thank Lucas van der Zee and Matteo Barberis for insightful discussions. This work was financially supported by various grants to H.V.W., such as from the Netherlands Organization for Scientific Research (NWO) in the integrated program of WOTRO [W01.65.324.00/project 4] Science for Global Development as well as by various systems biology grants, including Synpol: EU-FP7 [KBBE.2012.3.4-02 #311815], Corbel: EU-H2020 [NFRADEV-4-2014-2015#654248], Epipredict: EU-H2020 MSCA-ITN-2014-ETN:

Marie Skłodowska-Curie Innovative Training Networks (ITN-ETN) [#642691], and BBSRC China: [BB/J020060/1].
Contributions. T.D.G.A.M. performed the computational analyses with assistance from V.O. and H.V.W. H.V.W. conceived and supervised the study, provided scientific leadership, and produced the mathematical proof, which was checked by T.D.G.A.M. S.N., S.R., H.V.W. and T.D.G.A.M. identified mistakes in the Geenen et al. results, which T.D.G.A.M. then corrected. T.D.G.A.M. and H.V.W. wrote the manuscript.

Ch. 7 - GEMMs

Acknowledgments. This work was supported in part by the research Priority Area of the University of Amsterdam. E.W-T was financed by a grant within Mobilność Plus V from the Polish Ministry of Science and Higher Education (Grant 1639/MOB/V/2017/0). We gratefully acknowledge Eugenie Troia and Hugo Pineda Hernández for their kind help with HPLC and Stefania Astrologo for pointing out the existing datasets. We thank Chiara Damiani for discussing possible extensions of the MaREA methodology. We are thankful to Jakub M. Tomczak for helping out at the final stage of preparing the manuscript. We thank Pernette Verschure for co-supervision of D.P.
Contributions. Conceptualization, E.W-T.; methodology, E.W-T.,T.D.M and H.V.W; software, T.D.M; experimental work, E.W-T. and D.P.; validation, T.D.M. and E.W-T.; formal analysis, E.W-T.,T.D.M and H.V.W; investigation, E.W-T., T.D.M and H.V.W; resources, T.D.M and E.W-T.; writing–original draft preparation, E.W-T.,T.D.M and H.V.W; writing–review and editing, E.W-T.,T.D.M and H.V.W; visualization, E.W-T. and T.D.M; supervision, H.V.W.; project administration, E.W-T.

References

[1] J. of Salisbury and D. D. McGarry. *The Metalogicon of John of Salisbury, a Twelfth-century Defense of the Verbal and Logical Arts of the Trivium. Translated, with an Introduction and Notes, by Daniel D. McGarry*. University of California Press, 1955.

List of publications

- M. Barberis, **T.D.G.A. Mondeel**, Unveiling Forkhead-mediated regulation of yeast cell cycle and metabolic networks, Comput. Struct. Biotechnol. J. (2022). In Press.
- E. Węglarz-Tomczak, **T.D.G.A. Mondeel**, D.G.E. Piebes, H. V. Westerhoff, Simultaneous Integration of Gene Expression and Nutrient Availability for Studying the Metabolism of Hepatocellular Carcinoma Cell Lines, Biomolecules. 11 (2021) 490. 10.3390/biom11040490.
- **T.D.G.A. Mondeel**, O. Ivanov, H. V. Westerhoff, W. Liebermeister, M. Barberis, Clb3-centered regulations are recurrent across distinct parameter regions in minimal autonomous cell cycle oscillator designs, Npj Syst. Biol. Appl. 6 (2020) 8. 10.1038/s41540-020-0125-0.
- A. N. Kolodkin, R.P. Sharma, A.M. Colangelo, A. Ignatenko, F. Martorana, D. Jennen, J.J. Briedé, N. Brady, M. Barberis, **T.D.G.A. Mondeel**, M. Papa, V. Kumar, B. Peters, A. Skupin, L. Alberghina, R. Balling, H. V. Westerhoff, ROS networks: designs, aging, Parkinson's disease and precision therapies, Npj Syst. Biol. Appl. 6 (2020) 34. 10.1038/s41540-020-00150-w.
- **T.D.G.A. Mondeel**, P. Holland, J. Nielsen, M. Barberis, ChIP-exo analysis highlights Fkh1 and Fkh2 transcription factors as hubs that integrate multi-scale networks in budding yeast, Nucleic Acids Res. (2019) 1–17. 10.1093/nar/gkz603.
- Holland, P., Nielsen, J., **Mondeel, T. D. G. A.**, & Barberis, M. (2019). Coupling Cell Division to Metabolic Pathways Through Transcription. In Encyclopedia of Bioinformatics and Computational Biology (Vol. 3, pp. 74–93). Elsevier. 10.1016/B978-0-12-809633-8.20081-2.
- **Mondeel, T. D. G. A.**, Crémazy, F., & Barberis, M. (2018). GEMMER: GEnome-wide tool for Multi-scale Modeling data Extraction and Representation for Saccharomyces cerevisiae. Bioinformatics, 34(12), 2147–2149. 10.1093/bioinformatics/bty052.
- **Mondeel, T. D. G. A.**, Astrologo, S., Zhang, Y., & Westerhoff, H. V. (2018). NET works after all? Engineering robustness through diversity. IFAC-PapersOnLine, 51(19), 128–137. 10.1016/j.ifacol.2018.09.007.
- **Mondeel, T. D. G. A.**, Ogundipe, V., & Westerhoff, H. V. (2018). [Re] Predicting metabolic biomarkers of human inborn errors of metabolism. ReScience, 4(1), 1–12. 10.5281/zenodo.1254630.
- Abudukelimu, A., **Mondeel, T. D. G. A.**, Barberis, M., & Westerhoff, H. V. (2017). Learning to read and write in evolution: from static pseudoenzymes and pseudosignalers to dynamic gear shifters. Biochemical Society Transactions, 45(3), 635–652. 10.1042/BST20160281.
- **Mondeel, T. D. G. A.**, Rehman, S., Zhang, Y., Verma, M., Dürre, P., Barberis, M., & Westerhoff, H. V. (2016). Maps for when the living gets tough: Maneuvering through a hostile energy landscape. IFAC-PapersOnLine, 49(26), 364–370. 10.1016/j.ifacol.2017.03.002

- Westerhoff, H. V., Nakayama, S., **Mondeel, T. D. G. A.**, & Barberis, M. (2015). Systems Pharmacology: An opinion on how to turn the impossible into grand challenges. Drug Discovery Today: Technologies, 15, 23–31. 10.1016/j.ddtec.2015.06.006.