CHALMERS UNIVERSITY OF TECHNOLOGY
Gothenburg, Sweden
www.chalmers.se

Archeologists have revealed traces of intentional fermentation of grains which can be dated back to 10,000 – 12,000 years ago, coinciding or even preceding the estimated origins of agriculture and sedentary societies. This suggests that fermentation technologies, such as bread making and production of alcoholic beverages, have been a crucial activity in human history, since its earliest days.

Saccharomyces cerevisiae, a unicellular organism also known as baker's yeast, has provided a natural platform to drive fermentation processes, mainly due to its capacity to ferment sugars into ethanol. The transformation of nutrients, such as sugars, inside cells generates energy and the necessary precursors that the cells need for survival and growth. This process is known as metabolism, and is composed of thousands of chemical reactions. These reactions require enzymes and proteins, encoded by the genes of cells.

Throughout history humans have learned to use the metabolic potential of *Saccharomyces cerevisiae* and other yeasts to obtain chemical products that can be beneficial for society. This has resulted in development of applications for production of widely used pharmaceuticals, including insulin and artemisinic acid, flavors, fragrances, cosmetics, and fuel precursors.

In this thesis I use different quantitative approaches in systems biology to understand how different yeast species have evolved proteins that enable them to adapt to diverse environmental conditions. Furthermore, mathematical models and software resources were developed to aid to understand how the differences between enzymes affect the function of the cell. These models were used to predict how yeast cells can be engineered to increase their production of desired products, offering a successful example on an increased production of heme inside of *S. cerevisiae* cells. Heme is an important precursor of the protein that carries oxygen in human blood.

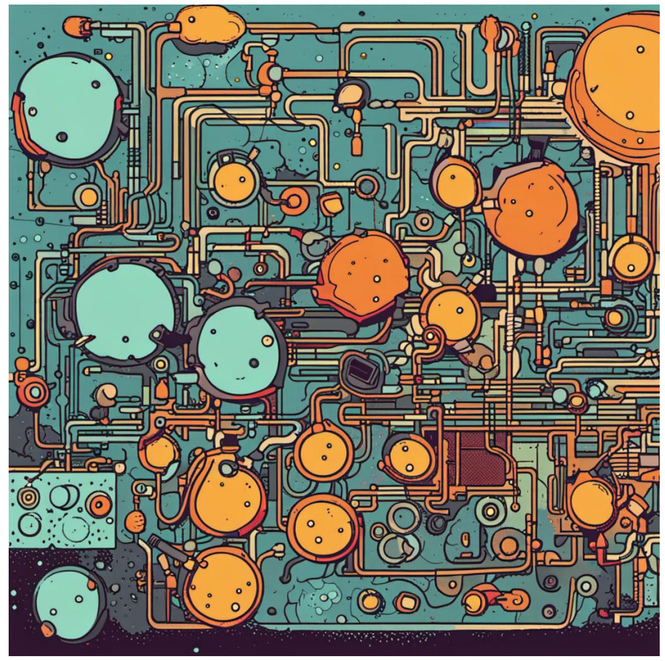
Finally, this thesis provides the community of biological scientists with models, software tools and methods for gaining understanding of the role of enzymes in living systems, and how they can be used for directed engineering purposes, such as the production of chemicals and pharmaceuticals; but also, for gaining basic knowledge on how cells function and interact with their environment, which has the potential to contribute to our understanding of human disease.

CHALMERS
UNIVERSITY OF TECHNOLOGY

Iván Domenzain del Castillo Cerecer A systems biology understanding of protrein constraints in the metabolism of budding yeasts. 2023

CHALMERS
UNIVERSITY OF TECHNOLOGY

PHD THESIS



A systems biology understanding of protein constraints in the metabolism of budding yeasts

Iván Domenzain del Castillo Cerecer

DEPARTMENT OF LIFE SCIENCES
CHALMERS UNIVERSITY OF TECHNOLOGY
Gothenburg, Sweden 2023
www.chalmers.se

A systems biology understanding of protein constraints in the metabolism of budding yeasts

IVÁN DOMENZAIN DEL CASTILLO CERECER

CHALMERS
UNIVERSITY OF TECHNOLOGY

Division of Systems and Synthetic Biology

Department of Life Sciences

CHALMERS UNIVERSITY OF TECHNOLOGY

Gothenburg, Sweden 2023

A systems biology understanding of protein constraints in the metabolism of budding yeasts

IVÁN DOMENZAIN DEL CASTILLO CERECER
ISBN 978-91-7905-911-8

Doktorsavhandlingar vid Chalmers tekniska högskola
Ny serie nr 5377
ISSN 0346-718X

Division of Systems and Synthetic Biology
Department of Life Sciences
Chalmers University of Technology
SE-412 96 Gothenburg
Sweden
Telephone + 46 (0)31-772 1000

Cover illustration:
”Modular synthesis, modular life” Created using NightCafe studio using the prompt: “*Draw a cartoon of a yeast cell being controlled by a modular synthesizer, the creation should express the inter connectivity and complexity of cellular processes using the analogy of complex connections in a synthesizer*"

Printed by Chalmers Digitaltryck
Gothenburg, Sweden 2023

To the memory of Julieta, Cuco and Uriel, with all my love

A systems biology understanding of protein constraints in the metabolism of budding yeasts

Iván Domenzain del Castillo Cerecer
Department of Life Sciences
Chalmers University of Technology

Abstract

Fermentation technologies, such as bread making and production of alcoholic beverages, have been crucial for development of humanity throughout history. *Saccharomyces cerevisiae* provides a natural platform for this, due to its capability to transform sugars into ethanol. This, and other yeasts, are now used for production of pharmaceuticals, including insulin and artemisinic acid, flavors, fragrances, nutraceuticals, and fuel precursors. In this thesis, different systems biology methods were developed to study interactions between metabolism, enzymatic capabilities, and regulation of gene expression in budding yeasts.

In **paper I**, a study of three different yeast species (*S. cerevisiae, Yarrowia lipolytica* and *Kluyveromyces marxianus*), exposed to multiple conditions, was carried out to understand their adaptation to environmental stress. **Paper II** revises the use of genome-scale metabolic models (GEMs) for the study and directed engineering of diverse yeast species. Additionally, 45 GEMs for different yeasts were collected, analyzed, and tested. In **paper III**, GECKO 2.0, a toolbox for integration of enzymatic constraints and proteomics data into GEMs, was developed and used for reconstruction of enzyme-constrained models (ecGEMs) for three yeast species and model organisms. Proteomics data and ecGEMs were used to further characterize the impact of environmental stress over metabolism of budding yeasts.

On **paper IV**, gene engineering targets for increased accumulation of heme in *S. cerevisiae* cells were predicted with an ecGEM. Predictions were experimentally validated, yielding a 70-fold increase in intracellular heme. The prediction method was systematized and applied to the production of 102 chemicals in *S. cerevisiae* (**Paper V**). Results highlighted general principles for systems metabolic engineering and enabled understanding of the role of protein limitations in bio-based chemical production. **Paper VI** presents a hybrid model integrating an enzyme-constrained metabolic network, coupled to a gene regulatory model of nutrient-sensing mechanisms in *S. cerevisiae*. This model improves prediction of protein expression patterns while providing a rational connection between metabolism and the use of nutrients from the environment.

This thesis demonstrates that integration of multiple systems biology approaches is valuable for understanding the connection of cell physiology at different levels, and provides tools for directed engineering of cells for the benefit of society.

Keywords: stress adaptation; metabolism; omics analysis; enzyme capacity; genome-scale modeling; metabolic engineering; gene regulation; systems biology

List of publications

This thesis is based on the following publications and manuscripts:

Paper I: Stress-induced expression is enriched for evolutionarily young genes in diverse budding yeasts.
Doughty TW, Domenzain I, Millan-Oropeza A et al. Nat Commun 2020, DOI: 10.1038/s41467-020-16073-3.

Paper II: Evaluating accessibility, usability and interoperability of genome-scale metabolic models for diverse yeasts species.
Domenzain, I., Li, F., Kerkhoven, E. J., & Siewers, V. (2021). FEMS Yeast Research, 21(1). https://doi.org/10.1093/femsyr/foab002

Paper III: Reconstruction of a catalogue of genome-scale metabolic models with enzymatic constraints using GECKO 2.0.
Domenzain, I., Sánchez, B., Anton, M., Kerkhoven, E. J., Millán-Oropeza, A., Henry, C., Siewers, V., Morrisey, J. P., Sonnenschein, N. and Nielsen, J. (2022). Nature Communications, 13(1), 3766. https://doi.org/10.1038/s41467-022-31421-1

Paper IV: Genome-scale modeling drives 70-fold improvement of intracellular heme production in Saccharomyces cerevisiae.
Ishchuk, O. P., Domenzain, I., Sánchez, B. J., Muñiz-Paredes, F., Martínez, J. L., Nielsen, J., & Petranovic, D. (2022). Proceedings of the National Academy of Sciences, 119(30), e2108245119. https://doi.org/10.1073/pnas.2108245119

Paper V: Computational biology predicts metabolic engineering targets for increased production of 102 valuable chemicals in yeast.
Domenzain, I., Lu, Y., Shi, J., Lu, H., & Nielsen, J. (2023). BioRxiv. https://doi.org/10.1101/2023.01.31.526512

Paper VI: A novel yeast hybrid modeling framework integrating Boolean and enzyme-constrained networks enables exploration of the interplay between signaling and metabolism.
Österberg, L., Domenzain, I., Münch, J., Nielsen, J., Hohmann, S., & Cvijovic, M. (2021). PLOS Computational Biology, 17(4), e1008891. https://doi.org/10.1371/journal.pcbi.1008891

Additional papers and manuscripts not included in this thesis:

Paper VII: RAVEN 2.0: A versatile toolbox for metabolic network reconstruction and a case study on Streptomyces coelicolor.
Wang, H., Marcišauskas, S., Sánchez, B. J., Domenzain, I., Hermansson, D., Agren, R., Nielsen, J. and Kerkhoven, E. J. (2018). *PLOS Computational Biology*, *14*(10), e1006541. https://doi.org/10.1371/journal.pcbi.1006541

Paper VIII: A consensus S. cerevisiae metabolic model Yeast8 and its ecosystem for comprehensively probing cellular metabolism.
Lu, H., Li, F., Sánchez, B. J., Zhu, Z., Li, G., Domenzain, I., Marcisauskas, S., Anton, P. M., Lappa, D., Lieven, C., Beber, M. E., Sonnenschein, N., Kerkhoven, E. J. and Nielsen, J. (2019). Nature Communications. https://doi.org/10.1038/s41467-019-11581-3

Paper IX: An atlas of human metabolism.
Robinson, J. L., Kocabaş, P., Wang, H., Cholley, P. E., Cook, D., Nilsson, A., Anton, M., Ferreira, R., Domenzain, I., Billa, V., Limeta, A., Hedin, A., Gustafsson, J., Kerkhoven, E. J., Svensson, L. T., Palsson, B. O., Mardinoglu, A., Hansson, L., Uhlén, M. amd Nielsen, J. (2020). Science Signaling. https://doi.org/10.1126/scisignal.aaz1482

Paper X: Benchmarking accuracy and precision of intensity-based absolute quantification of protein abundances in Saccharomyces cerevisiae.
Sánchez, B. J., Lahtvee, P. J., Campbell, K., Kasvandik, S., Yu, R., Domenzain, I., Zelezniak, A., and Nielsen, J. (2021). Proteomics, 21(6). https://doi.org/10.1002/pmic.202000093

Paper XI: Yeast metabolic innovations emerged via expanded metabolic network and gene positive selection.
Lu, H., Li, F., Yuan, L., Domenzain, I., Yu, R., Wang, H., Li, G., Chen, Y., Ji, B., Kerkhoven, E. J. and Nielsen, J. (2021).. Molecular Systems Biology, 17(10). https://doi.org/10.15252/msb.202110427

Paper XII: Identification of a novel gene required for competitive growth at high temperature in the thermotolerant yeast Kluyveromyces marxianus.
Montini, N., Doughty, T. W., Domenzain, I., Fenton, D. A., Baranov, P. V, Harrington, R., Nielsen, J., Siewers, V. and Morrissey, J. P. (2022). Microbiology, 168(3). https://doi.org/https://doi.org/10.1099/mic.0.001148

Paper XIII.- Improving the production of biologicals in Saccharomyces cerevisiae by overexpressing native target genes predicted by two proteome constrained genome-scale models
Veronica Gast, Feiran Li, Iván Domenzain, Mikael Molin and Verena Siewers. Manuscript.

Paper XIV: standard-GEM: standardization of open-source genome-scale metabolic models.
Mihail Anton, Eivind Almaas, Rui Benfeitas, Sara Benito-Vaquerizo, Lars M. Blanck, Iván Domenzain, Andreas Dräger, John M. Hancock, Cheewin Kittikunapong, Matthias König, Feiran Li, Ulf W. Liebal, Hongzhing Lu, Hongwu Ma, Radhakrishnan Mahadevan, Costas Maranas, Adil Mardinoglu, Jens Nielsen, Juan Nogales, Marco Pagni, Jason A. Papin, Kiran Raosaheb Patil, Nathan D. Price,

Jonathan L. Robinson, Benjamín J. Sánchez, María Suárez-Diez, Snorre Sulheim, L. Thomas Svensson, Bas Teusink, Wanwipa Vongsangnak, Hao Wang, Ahmad A. Zeidan, Eduard J. Kerkhoven. bioRxiv 2023.02.21.512712. Manuscript.

Paper XV: Model-driven polyols production via oxidoreductive pathway in *Candida intermedia*.
Kameshwara V. R. Peri, Iván Domenzain, Abril Valverde, Fábio Faria Oliveira, Jens Nielsen and Cecilia Geijer.
Manuscript

Paper XVI: The metabolic landscape of the eukaryotic cell cycle.
Iván Domenzain, Kate Campbell, Jens Nielsen. Manuscript

Paper XVII: The control of metabolic flux revisited.
Iván Domenzain, Yu Chen, Jens Nielsen. Manuscript.

Paper XVIII: Model-driven construction of platform strains for increased production of bioactive molecules in yeast.
Ricardo Bisquert, Iván Domenzain, Jens Nielsen and José Manuel Guillamón.
Manuscript.

Paper XIX: Reconstruction, simulation and analysis of enzyme-constrained metabolic models using GECKO toolbox 3.0.
Yu Chen , Johan Gustafsson , Albert Tafur Rangel , Mihail Anton , Iván Domenzain, Cheewin Kittikunapong , Feiran Li , Le Yuan , Jens Nielsen.
Manuscript.

Contribution summary

Paper I.- Designed the software for integrated analysis of transcriptomics and proteomics across conditions and organisms. Analyzed transcriptomics and proteomics and gene orthology data. Contributed to discussion of results, writing and editing of the final manuscript.

Paper II.- Performed literature review, conceived the model evaluation framework, creation of the software testing pipeline. Analyzed the data and co-wrote the manuscript.

Paper III.- Conceived the main ideas leading to the extension of GECKO and implemented them into the software, designed the methodology for the application cases, analyzed proteomics data and simulation results. Wrote the manuscript.

Paper IV.- Developed the method and software for prediction of gene targets with enzyme-constrained models. Analyzed simulation results and contributed to write the original manuscript draft and editing.

Paper V.- Conceived the prediction method, developed the software, analyzed results, and wrote the original manuscript.

Paper VI.- Designed the methods and software for incorporating the metabolic and enzyme-constrained network into the hybrid modeling. Curated and analyzed proteomics data, analyzed, and discussed simulation results and co-wrote the original manuscript.

Preface

This dissertation serves as partial fulfillment of the requirements to obtain the degree of Doctor of Philosophy at the Department of Life sciences at Chalmers University of Technology. The PhD studies were carried out between March 2017 and September 2023 at the Division of Systems and Synthetic Biology (SysBio) under the supervision of Jens Nielsen. The project was co-supervised by Verena Siewers and Eduard J. Kerkhoven and examined by Ivan Mijakovic. The project was funded by the European Union's Horizon 2020 research and innovation program projects CHASSY (grant no. 720824) and DD-decaf (grant no. 686070).

Iván Domenzain
September 2023

Table of contents

Abbreviations and symbols

ADP.- Adenosine diphosphate
ATP.- Adenosine triphosphate
CO_2.- Carbon dioxide
DNA.- Deoxyribonucleic acid
FBA.- Flux balance analysis
FVA.- Flux variability analysis
GEM.- Genome-scale metabolic model
gDw.- Grams of cell dry weight
ecGEM.- enzyme-constrained metabolic model
FDR.- False discovery rate
MCA.- Metabolic control analysis
NADH.- Nicotinamide adenine dinucleotide
NADPH.- Nicotinamide adenine dinucleotide phosphate
RNA.- Ribonucleic acid
TCA.- Tricarboxylic acids

1.- Background

This chapter has the intention of providing the reader with an understanding of the central subjects of the research presented in this thesis, namely, the relevance of the use and study of diverse yeasts for the humankind; systems biology as a scientific field; and how can the tools of systems biology be used for advancing our knowledge and manipulation of yeasts for biotechnological and medical uses.

I acknowledge systems biology as a rapidly evolving discipline, where novel methods and concepts are being developed frequently. Therefore, I have decided to present this section following a historical narrative, which describes the non-linear path followed by scientists in establishing and practicing a discipline. History is defined and told backwards in time, therefore, the selection of references and time points in the narrative, represent a snapshot of the ideas that have influenced those presented in this thesis.

A historical narrative facilitates showing that science is not a well-defined set of rules, theories, and methods that operates in a one-directional way, according to a given program. But rather, I see science as an evolving complex network of individuals, institutions, infrastructures, concepts, and methods, explored by human curiosity, sometimes driven by necessity, but many others just by our grit and insatiable wish for understanding.

1.1 Why to write a Doctoral thesis on yeasts?

Around 10,000 years ago humans started the development of agriculture, which led to an unprecedented availability and accumulation of food resources that fostered the development of the first sedentary societies and organized civilizations, setting the basis for a never-ending history of cultural innovations. Archeologists have revealed that vestiges of intentional fermentation of grains can be dated back to around the same era, some of them even dating it to 13,000 years ago[1,2]. This suggests that fermentation technologies have accompanied human history since its origins. Production of alcoholic beverages and bread making are fermentation-based processes, ubiquitous across eras and cultures, made possible by diverse yeast species that break down the sugars in nutrient rich foods into alcohol or organic acids. Even though yeast cells were first observed under a microscope by Anton Van Leeuwenhoek in 1680, understanding of yeasts as causal agents of fermentation, took shape over the next 200 years, involving notable scientists such as Lavoisier, Gay-Lussac, Louis Pasteur and Emil C. Hansen[3].

Yeasts are unicellular organisms, such as bacteria, however, their cells contain a well-defined nucleus, that stores their genetic material, and other intracellular compartments specialized in vital processes. Therefore, yeasts are classified as Eukaryote organisms (like plants, mushrooms, insects, reptiles, birds, humans, etc.),

within the Fungi kingdom and the Ascomycota phylum. Most yeasts that have been historically used by humans display asexual reproduction, in which a new organism develops from a bud due to cell division of an original cell, thus, they have been popularly called budding yeasts. In 1837 the term *Saccharomyces*, meaning "sugar fungus" in Latinized Greek, was coined by Julius Meyen to name budding yeasts[3]. These yeasts have been further classified into the *Saccharomycetales* order in more recent times[4].

Saccharomyces cerevisiae, widely known as baker's yeast, has been the most utilized and studied of all budding yeast species. Being a unicellular Eukaryote, with a millenary relation to humans, makes it a suitable platform for the study of cellular processes of signaling, division, aging, and death, also present in organisms as complex as the human body. Hence, its use in biological and medical sciences as a model organism has resulted in several Nobel prizes, mostly in the category of physiology and medicine. Examples of these include the discovery of DNA synthesis and replication (Kornberg, awarded in 1959); discovery of restriction enzymes (Arber, Nathans and Smith, in 1978); discovery of key regulators of the cell cycle (Hartwell, Hunt and Nurse in 2001); discovery of the protection of chromosomes by telomeres (Blackburn, Greider and Szostak in 2009); and discovery of mechanisms involved in autophagy (Ohsumi, in 2016)[5].

Furthermore, *S. cerevisiae* has been studied and engineered for its use in a wider variety of industrial purposes, spanning from the production of insulin, opioids, and other relevant pharmaceuticals for improving human health, as well as for production of chemicals, flavors, fragrances and cosmetics[6]. In more recent years, other non-conventional yeast species have attracted attention due to their unique phenotypic characteristics that can be leveraged in the industry, for instance, high accumulation of lipids in *Rhodotorula toruloides* and *Yarrowia lipolytica* for production of biofuel precursors; high temperature tolerance (up to 52°C) in *Kluyveromyces marxianus*[7]; and high tolerance towards acetic acid in *Zygosaccharomyces bailii*[8]. Overall, this shows the value of using yeasts and fermentation technologies for providing viable alternatives to conventional chemical processes, facilitating transition towards sustainable production frameworks.

The history of yeast domestication might be much more complex than it appears, resembling more a process of co-evolution, in which diverse yeasts have been adapted to different environments with different production purposes across thousands of years; whilst humans have been able to utilize yeasts for increasing nutritional value of food, production of alcohol, present in ritual practices across many cultures, and even for understanding biological processes that are present in our own cells.

1.2 Cellular metabolism

Despite their microscopic size, yeasts were first noticed by humans due to their effects over grains and fruits, namely, production of "air bubbles" and development of alcoholic and sour tastes. These observable phenomena are a result of the conjunction of myriads of biochemical transformations (i.e., reactions) that take place inside of cells. These reactions operate in a network manner, initially taking nutrients from the environment as substrates, and converting them into molecular building blocks for cellular growth and byproducts, such as CO_2 and ethanol (**figure 1A**). The nutrients demanded by cells can be classified into a carbon source, energy source, nitrogen source, minerals, and vitamins. It is often that the carbon and energy source are the same compound, being glucose the most common across organisms[9].

In this network of reactions, energy (in the form of Gibbs free energy, a measure of the potential of chemical transformation in a substance) is extracted from the substrate and stored as an energy currency molecule (i.e., adenosine triphosphate, ATP) and other cofactors. This forms the basis of the process known as catabolism, which operates in coordination with anabolism, the assembly of the produced building blocks into macromolecular cellular components (carbohydrates, lipids, proteins and nucleic acids), powered by the energy stored as ATP[10]. Together, catabolism and anabolism form cellular metabolism, illustrated in **figure 1B**.

Metabolism encompasses thousands of biochemical reactions and hundreds of different compounds, referred to as metabolites[11]. Its basic structure, conserved across all living forms, it is at the core of cellular function as it enables all the other processes that make life to happen, such as self-maintenance, signaling, growth and reproduction[12].

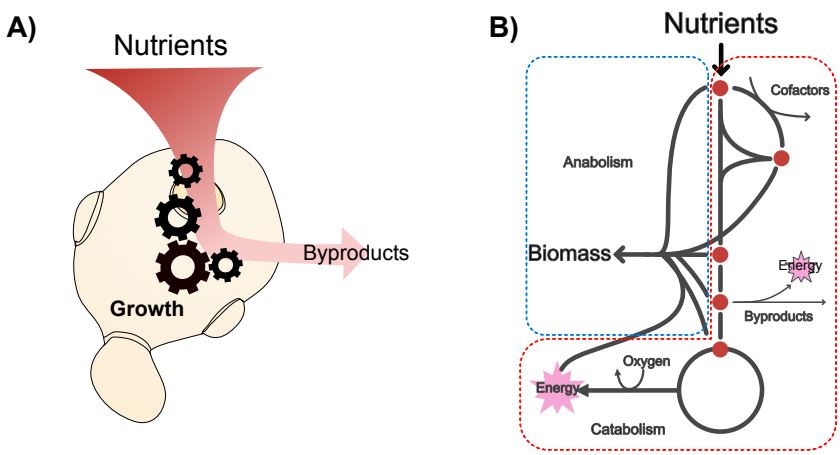


Figure 1.- A summarized view of cellular metabolism. A) Nutrients are broken down into byproducts, while extracting enough energy and components to run the cellular machinery

that enables growth. B) A more detailed view of the cellular machinery shows that nutrients are converted into byproducts, cofactors, energy, and molecular building blocks (catabolism, surrounded by a red dotted-line), then a series of mechanisms use the building blocks, energy and cofactors to generate new biomass or cellular growth (anabolism, surrounded by the blue dotted-line).

To facilitate understanding, metabolism is usually divided into different pathways in textbook and biochemistry literature. Each of these pathways has a particular structure and fulfills different purposes. The most studied metabolic pathways are those in central carbon and energy metabolism as they are present in all known kinds of living cells. A brief overview of these pathways is presented next.

Glycolysis, also known as the Embden-Meyerhof-Parnas pathway, is the pathway by which most living cells generate the necessary energy for growth from sugars. It consists of a series of 10 reactions divided into an "investment" part, in which ATP is required to phosphorylate or "activate" the substrate, and a yield section, in which a surplus of ATP is obtained from chemical transformations[13]. The final products are a net total of 2 ATP molecules and 2 molecules of pyruvate per each molecule of glucose that enters the pathway. Pyruvate can either be fermented or undergo further modifications in the tricarboxylic acids cycle. A detailed review of the "design" principles behind each reaction step, emphasizing the constraints (thermodynamic or kinetic) limiting each of them was published by Bar-Even and collaborators in 2012 and can be found elsewhere[14].

Fermentative metabolism pathways are ubiquitous in nature, as they provide cells with a mechanism for restoring homeostasis, even in anaerobic environments. During the transformation of glucose into pyruvate in glycolysis, electrons are transferred from metabolite intermediates to cofactor molecules or "electron carriers". The most used electron carrier in catabolism is nicotinamide adenine dinucleotide (NADH, with its oxidized form NAD+). Glycolysis produces two moles of NADH per mole of catabolized glucose, but as the redox state of the cell is highly regulated, these NADH molecules need to be restored into their oxidized form. Thus, fermentation pathways reduce the resulting pyruvate into byproducts such as lactic acid, common in lactic acid bacteria and mammalian muscle cells, or ethanol, commonly produced by different budding yeast species.

The pentose phosphate pathway is a complex network of reactions that serves for multiple purposes. Its entry point is glucose-6-phosphate, the first metabolic intermediate in glycolysis, which is then oxidized into ribulose-5-phosphate plus a molecule of CO_2, in these first reactions 2 moles of NADPH are formed, which provide the reducing power necessary for anabolic processes like protein synthesis. Further down, 2 molecules of ribulose-5-phosphate can be transformed into the pentoses ribose-5-phosphate (R5P), essential for DNA and RNA synthesis, and xylulose-5-phosphate. The carbon atoms in these molecules can then be rearranged and transformed back into glycolysis intermediates to maximize energy production,

or further transformed into erythrose-4-phosphate (E4P), an essential building block for synthesis of aromatic amino acids[9].

The tricarboxylic acids cycle (TCA cycle) is a highly versatile pathway, brilliantly elucidated by the works of three Nobel laureates Otto Warburg, Albert Szent-Györgyi and Adolf Krebs[15]. In a simplified way, and assuming respiratory conditions with no loss of intermediates, this pathway takes the three carbons compound pyruvate, end product of glycolysis, and by a series of oxidative reactions transforms it into three molecules of CO_2, five redox packages (4 NADH and one $FADH_2$) and one energy currency molecule (GTP)[9]. The five redox packages can be further used to fuel cellular respiration to produce additional energy, in which the carried electrons are transferred to molecular oxygen, the final electron acceptor. Moreover, several of its intermediate metabolites (2-oxoglutarate, succinate and oxaloacetate) constitute fundamental building blocks of the cell and play major roles in amino acid biosynthesis[16].

Notably, the TCA cycle serves as an engine that adjusts the rate and direction of its revolutions according to the conditions that the cell is exposed to. In anaerobic conditions, the cycle operates for production of building blocks, but it is also likely to operate in both directions in order to balance redox potential. When glucose is not available and non-fermentable compounds constitute the main carbon source of the cell (e.g. ethanol or acetate), the TCA cycle, aided by shunt reactions, facilitates production of phosphoenolpyruvate (PEP), a glycolytic intermediate upstream pyruvate, enabling a reversed glycolytic process (gluconeogenesis) in which glucose-6-phosphate and other building blocks can be produced by the cell. The 12 main precursor metabolites for biomass synthesis, the pathways in which they are mostly produced, and the biosynthetic pathways that they participate in, are summarized in **table 1**.

The electrons carried by the redox packages formed in a clockwise turn of the TCA cycle can be leveraged for production of additional ATP by the cell. Mechanisms for creating a current of electrons through different membrane proteins, and using this work for pumping protons from the inside to the outside of such membrane against a concentration gradient, are present across all domains of life[17]. The flow of electrons is taken at the end of the chain of pumping proteins by an acceptor molecule, oxygen being the one used by yeasts, animals, and many bacteria.

Prokaryote cells carry out this process in their cell membranes, pumping protons to the immediate surroundings, whilst yeast and all eukaryotes carry this out in the intermembrane space of specialized compartments with inner folded membranes, the mitochondria. Flow of protons from the exterior to the interior of the space enclosed by these membranes, activates the rotational motion of a specialized complex of proteins, ATP synthase, which consists of stator and a rotor part. The latter enables phosphorylation of ADP molecules to ATP while it spins[18], resembling the function of water turbine in a hydroelectric dam.

This additional energy producing series of mechanisms is known as oxidative phosphorylation, in opposition to glycolysis, known as substrate-level phosphorylation. The elucidation of this pathway took years of relentless, and controversial, work by Peter Mitchell and Jennifer Moyle. The former was awarded the Nobel prize in 1978, for his development of the chemiosmotic theory of cell energetics, while Jennifer Moyle, as very often in science, remained unacknowledged for her meticulous molecular work and measurements which majorly supported the ideas and hypotheses of Mitchell[15]. This theory of cell energetics has been suggested to be one of the more counterintuitive postulates in biology, comparable to the level of the hypothesis of natural selection by Charles Darwin[17]. However the theory is now a well-accepted and fundamental part of the current paradigm of cell physiology in science[18].

Oxidative phosphorylation greatly increases the amount of ATP that can be generated by using the nutrients, in particular the carbon source, fed to the cell. Its thermodynamic efficiency has been estimated to be as high as 0.42, theoretically, which surpasses that of the most efficient hydrogen fuel cells made by humans[9], and also that of most thermal engines, which are the basis of current human industries.

Even more surprising is the scale and rates at which this process takes place inside of cells. Eukaryotes have evolved mitochondrial structures that offer enormous amounts of surface membrane area per volume. Extending the full membrane area of these in front of our eyes would cover as much as four full football pitches, all packed by electron transfer complexes and the ATP synthase engine[17]. Using the energy carried by the transfer of electrons enables proton pumping, up to a rate that can maintain an electric potential of 150-200 mV, across a membrane of 5-6 nm. Upscaling these numbers to a macroscopic world would imply that an electric potential of $30x10^6$ V/m is generated by the electron transport chain[19]. In human experience, this would be equivalent to a person been "hit" by a thunderbolt.

Consequently, this massive electric potential causes a flow of protons back again into the mitochondria across the inner channels of ATP synthase, whose rotor can spin up to 100 revolutions per seconds, generating several molecules of ATP per rotation in each of its copies[20]. No wonder why mitochondria are often called "the powerhouse of cells".

Table 1.- Twelve main precursor metabolites for cellular biosynthesis.

Precursor metabolite	**Pathway**	**Biosynthetic associated processes**
Glucose-6-phosphate	Glycolysis/gluconeogenesis	Carbohydrates production
Fructose-6-phosphate	Glycolysis/gluconeogenesis	Carbohydrates production
Glyceraldehyde-3-phosphate	Glycolysis/gluconeogenesis	Phospholipids biosynthesis

3-phosphoglycerate	Glycolysis/gluconeogenesis	Amino acid and nucleotides metabolism
Phosphoenolpyruvate	Glycolysis/gluconeogenesis	Synthesis of aromatic amino acids
Pyruvate	Glycolysis	Amino acids metabolism
Ribose-5-phosphate	Pentose phosphate pathway	Amino acid and nucleotides metabolism
Erythrose-4-phosphate	Pentose phosphate pathway	Synthesis of aromatic amino acids
Acetyl-CoA	TCA cycle	Lipids metabolism
2-Oxoglutarate	TCA cycle	Amino acids metabolism
Succinyl-CoA	TCA cycle	Protein biosynthesis
Oxaloacetate	TCA cycle	Amino acid and nucleotides metabolism

1.3 The central dogma of molecular biology

In 1943 Erwin Schrödinger, an Austrian physicist, delivered a series of lectures at Trinity College Dublin in Ireland, in which he addressed the question *What is life?* He followed thermodynamic arguments to conclude that life is fundamentally a state of organization in matter, which keeps itself under a higher level of molecular order in contrast to its surroundings, by means of boundaries that define the interior and exterior of the system.

The processes for maintaining this state of organization require a constant flow of energy and materials in and out the boundaries. At the times of Schrödinger, extensive studies on the mechanisms of inheritance of acquired characteristics by living organisms were available. But the fundamental pieces of inheritance mechanisms were lacking, therefore, he hypothesized that some sort of molecular information needed to be transferred from cells to their daughters, and then stored in them, so that the required processes for life and its reproduction could be executed again. Schrödinger, as a quantum physicist, saw the necessity of an aperiodic crystal as the molecular basis of this information storage and transfer. These lectures were later published as the essay *What is life?* in 1944[21].

Schrödinger ideas inspired a generation of crystallographers to investigate the structure of biological macromolecules using the principles of radiation and quantum chemistry and physics. Francis H. Crick, coming from a prominent direct academic lineage, holding the names of J. J. Thomson, Lord Rayleigh, W. H. Bragg, and Max

Perutz[22], worked together with James D. Watson and Rosalind Franklin in deciphering the identity and structure of such aperiodic crystal.

In 1953, Crick and Watson published a scientific article in *Nature*, entitled "Molecular structure of nucleic acids: A structure for Deoxyribose nucleic acid", in which they proposed the double helix structure of DNA, with nucleotide bases at its core, for which they found specific pairings (adenine – thymine and guanine – cytosine) that, according to them, "immediately suggests a possible copying mechanism for the genetic material"[23]. This was acknowledged as the missing piece in genetic inheritance mechanisms by the scientific community and Crick and Watson were awarded the Nobel prize in physiology and medicine, together with Maurice Wilkins, in 1962. Although their proposed structure was based on the experimental work of Rosalind Franklin, in particular "photo 51"(**figure 2**), she was excluded from the awarded prize, as many other notable women scientists have been unjustly omitted from the history of scientific discovery[24].

Later on, Francis Crick systematized the available knowledge on molecular biology in a couple of articles, in 1958[25] and 1970[26], setting the foundations of the well-known central dogma of molecular biology. This concise formulation of molecular biology consists of the following. First, proteins are the core components in cells that enable cellular processes to take place, due to its "enzymatic" or catalytic activity.

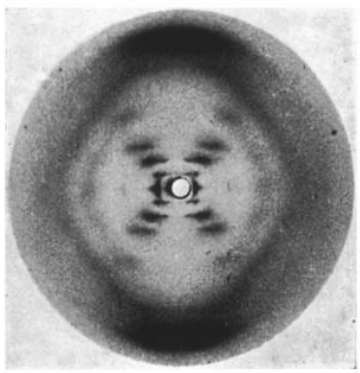


Figure 2.- Photo 51, X-ray crystallography image of DNA. Raymond Gosling. King's College London. 1952. Source: http://www-project.slac.stanford.edu/wis/images/photo_51.jpg.

Proteins consist of long sequences of different amino acids tied together by covalent bonds, there are 20 different amino acids that can be combined in particular sequences to form specific proteins. The information for synthesizing each protein is stored in DNA, the sequence of nucleotide bases at the core of its helix structure, organized by triplets, corresponds to a sequence of specific amino acids. Crick demonstrated the correspondence 4 bases –> organized in triplets –> 20 amino acids, using combinatorial and heuristic arguments.

DNA can generate copies of itself by replication and then be transcribed, in order to produce RNA, a single stranded molecule that keeps the sequence of nucleotide bases, containing the code for proteins. At the time of his articles, little was known about the role of ribosomes in protein synthesis, however, it was hypothesized that some sort of protein with catalytic activity could be responsible for taking adapted pieces of RNA and use them as a template for putting a sequence of amino acids together, thus, RNA is translated into a protein.

It must be mentioned that Francis Crick acknowledged the importance that post-translational modification of proteins may play on their final function, especially chemical modification of discrete amino acid sequences and 3D protein structure. Nonetheless, his reasoning prioritized the linear sequence of amino acids, and its relation to DNA code, as the central factor differentiating cellular function between organisms and shaping inheritance. A schematic view of the central dogma of molecular biology, aided by more modern knowledge (e.g., ribosomes as protein synthesizers) is shown in **figure 3**.
Central to the theory assembled by Crick and others of his time were the following points:

- The main function of proteins is to act as enzymes.
- "Once that information has passed into a protein *it cannot get out again*"[25]. The transfer from protein to protein or from protein to DNA is impossible. Information meaning the precise determination of sequences, either of bases in nucleic acids or amino acid residues in proteins.
- "The central dogma is intended to apply only to modern organisms, and not to events remote in the past, such as the origin of life or the origin of the code"[26].

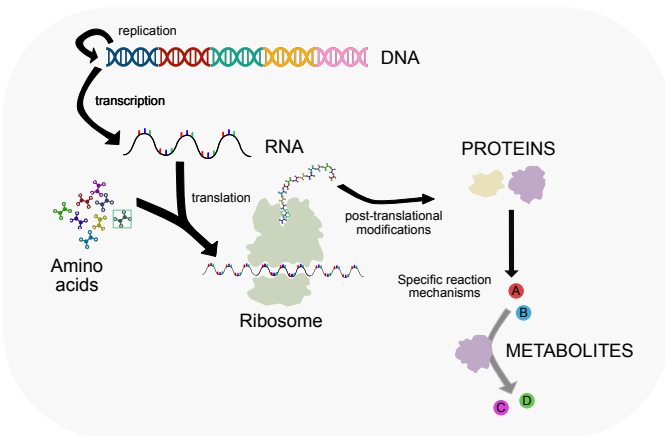


Figure 3.- The central dogma of molecular biology. DNA contains the genetic information to reproduce all the necessary processes for life, it reproduces itself by DNA repli[26]cation. DNA gets transcribed into RNA and messenger fragments of it are taken by ribosomes, that assemble sequences of amino acids, according to the code in RNA. Linear sequences of amino

acids undergo folding and post-translational modifications that convert them into functional proteins that enable chemical transformations and biological processes. Uppercase labels indicate macromolecular components of the cell. Lowercase labels indicate biological processes that transfer information from one molecular level to lower ones.

Nevertheless, it should also be mentioned that Francis Crick was aware of the limitations of these ideas. In his 1958 article, exposing the central dogma, he expressed the following words, displaying a radical Popperian approach to science[27], regarding the process of protein synthesis: "*some of these points are now ripe for a direct experimental attack*"[25], and even expressed that potential discovery of an information transfer not considered by the central dogma would "*shake the whole intellectual basis of molecular biology*"[25]. Hence, even though Crick and collaborators used the term "dogma" for their theory, they did not conceive it, neither communicated it, as a universal truth.

The discovery of DNA and the enunciation of the central dogma facilitated the acceptance of molecular biology as its own scientific discipline, differing from traditional biochemistry, and pure genetics, in its scope, questions, and methods. For molecular biology the correspondence gene – protein provides a foundation for the exploration of gene "function", relying on directed molecular experiments to test the relation genotype – phenotype as its main methodological tool (e.g., single gene knock-outs)[28].

The scope of molecular biology brought an ontologically reductionist perspective to biology, in which living organisms could be reduced to the study of the molecules constituting them, which by themselves could be explained by the principles of chemistry and, ultimately, physics. In this way, molecular biology offered a solution to the philosophical and methodological problems of 20th century biology[29].

1.4 What is Systems Biology?

According to Thomas S. Kuhn view of the scientific enterprise, once a scientific community have shaped methodological, ontological, and epistemological consensus through contrasting competing theories, a paradigm is established[30]. An established paradigm ensures scientific progress in a cumulative way, in which scientists focus on measuring all the missing quantities that would enable their theoretical framework to explain the observable phenomena that their field is concerned with.

No paradigm has proven to be flawless in the history of physico-chemical sciences[24], as observations that deviate from their conclusions arise often, or even internal contradictions can be found in them after detailed scrutiny. Therefore, "anomalies" accumulate during the lifetime of a paradigm, up to points in which the consensus around the paradigm cannot be held any longer.

In the case of molecular biology, such paradigmatic anomalies started to appear early in its development. One of this examples came 1960, when Denis Noble published a series of mathematical models of the heart's pacemaker mechanism that successfully captured experimental observations. It was by setting up a mutual dependency on a voltage potential at the cellular level, and ion transport activity at the protein level, that experimental observations on Purkinje fibers from sheep ventricles could be predicted[31,32]. This posed an example on a feedback mechanism between distant levels of biological organization, protein activity and cellular physiology, which should be, in principle according to the central dogma, dictated by lower-level mechanisms, like gene expression and protein activity itself.

The years spanning from 1950's – 70's saw a sudden rise of introduction of mathematical thinking and modeling in biological problems, based on principles of thermodynamics[33–36], reaction kinetics[37–41], and information theory[42–44]. This opened a wave of scientific development, that did not necessarily account for the central dogma, and provided predictive power. Common to all these lines of development was the reliance on a "systems thinking", rather than a "component thinking"[45], prevalent in molecular biology[46].

Systematization of the study of reaction kinetics in the context of metabolic networks in the 70's and 80's gave rise to the theory of metabolic control analysis (MCA)[37,47], its main prediction being that kinetic control of metabolic pathways did not reside in a single catalytic step, but it is rather distributed across several points in the network. This provides another example of the prediction of a network property from assembling detailed knowledge, at the molecular level, on a large scale.

In parallel, molecular biology developed further, especially in its methods. Genome sequencing became available in the 1970's[48,49], leading to an unprecedented accumulation of biological data and major efforts towards functional characterization of genes, following a gene – phenotypic function logic. During the 1980's several genome-wide sequencing projects were launched[50], and in the 1990's large-scale identification and characterization of genes was enabled by microarray technologies[51,52].

At the turn of the century, a full catalogue of thousands of associations gene – function across many different model organisms, from all domains of life, was made available by the gene ontology consortium[53]. Such an accumulation of new biological data fueled the development of detailed mathematical studies, relying on molecular knowledge, explaining biological phenomena at different spatial, temporal and organizational scales[54–60].

The emergence of a new discipline was fully realized with the foundation of two independent systems biology institutes, one in Seattle, USA, and the other in Tokyo, Japan, in the year 2000. Additionally, seminal publications by the founders of these institutes, Hiraoki Kitano[61,62] and Leroy Hood and Trey Ideker[63], provided a

framework to understand the scope, methods, tools, and questions of systems biology.

Kitano defined systems biology as a field that aims at understanding biological systems at system level, requiring a change in our notion of "what to look for" in biology. Based on detailed understanding of genes and proteins, the focus is on understanding system's structures, dynamics and control mechanisms, and also to develop strategies to modify and construct biological systems based on rational design principles[61,62]. This provides a high-level definition of the field, delimiting the epistemological subjects of systems biology (i.e., what counts as knowledge?).

On the other hand, Ideker's definition of systems biology argues that, in order to understand the relation between molecular interactions and global changes in biological systems, it is necessary to integrate various levels of global measurements together with mathematical models of the systems of interest. This should be guided by the following steps: define all the components of the system; perturb and monitor systems components systematically; reconcile experimental observations with model predictions; and design new perturbation experiments to distinguish between multiple model hypotheses[63]. In contrast to Kitano's definition, this is a more methodological definition of the field, highlighting the procedures and steps that count as valid for constructing systems biology knowledge.

From these definitions, two main aspects are of crucial importance at differentiating systems biology from other fields in biology: 1) understand complex biological organization and processes using detailed knowledge and measurements of molecular components, and 2) the use of mathematical models for understanding relations in between molecular components and predict systems responses. Albeit these epistemic and methodological differentiation, systems biology shares concerns that are central to other branches of biology, such as, information transfer, with molecular biology; characterization of adaptive states of cells and organisms, with physiology; definition of the succession of adaptive states, with developmental biology; and the appreciation that all aspects in an organism are products of selection, with ecology and evolutionary biology[64].

Denis Noble has been another major contributor to the definition and understanding of systems biology, not just with his science but also with his philosophical publications and dissertations. A fundamental idea across his works is the identification of an ontological difference between molecular biology and systems biology. Noble describes the ontology of molecular biology as a reductionist causal chain, in which genes play the most fundamental role and their code governs the processes taking place at higher levels of organization (upwards causation). In contrast, systems biology rejects this idea by acknowledging that transmission of information is not just one way in biological systems (downwards causation), and that functionality is something that can just be understood at the organism level, encompassing all the lower levels of organization[65–67]. The differences between these

opposite views are summarized in **figure 3** and explained in a cellular context in **figure 4**.

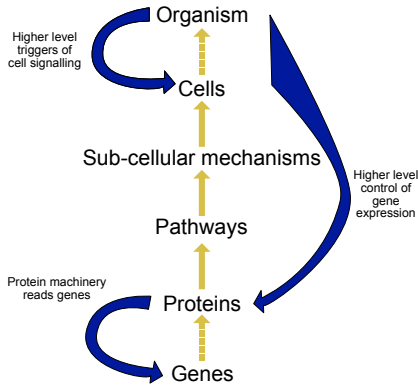


Figure 3.- Upwards and downwards causation in systems biology. Dark yellow arrows represent the upwards transfer of information, central to reductionist approaches. Blue arrows represent downwards causation or transfer of information introduced by systems biology. Reproduced from Noble, D., 2008[65].

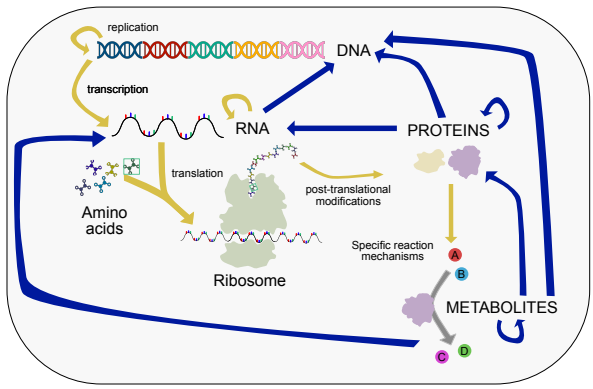


Figure 4.- Upwards and downwards causation in the context of a cell. Downwards information transfer between macromolecules (i.e., from higher to lower levels of organization, represented with blue arrows) complement the one-way information transfer of molecular biology (upwards transfer, represented by dark yellow arrows) and enable study of emergent system properties.

Bernhard Ø. Palsson defines another distinction in between different levels of causation in biological systems, which he has called the dual causation of systems biology[68]. In this way, physical laws and changes on the environment impose constraints during the lifespan of an organism. In order to survive, organisms adapt to these constraints, through phenotypic adaptations or proximal causation (e.g., transcriptional responses to environmental stress). Additionally, changes in the

genotype, at the level of sequence and composition, generate diverse populations of individuals, the most fit ones proliferate in the phenotypic landscape (distal causation), giving rise to continuous iterations of the dual causation cycle (proximal – distal), as depicted in **figure 5**.

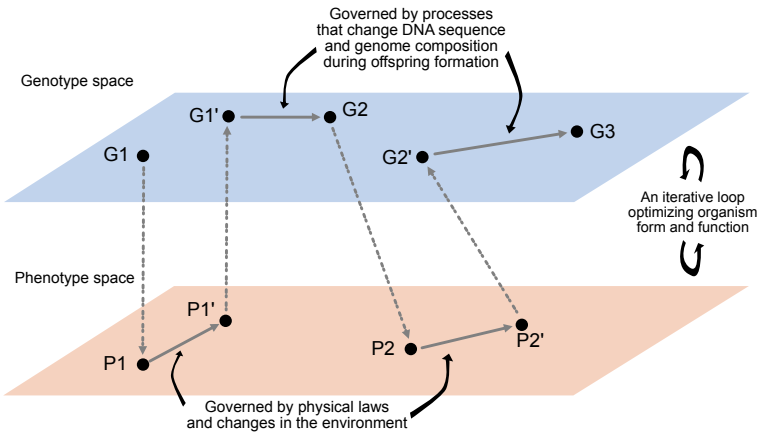


Figure 5.- Dual causation of the genotype-phenotype relation proposed by Bernhard Palsson. Adapted from Palsson, B. O., 2015, pp. 253[68].

1.5 Methodological approaches in systems biology

In the last two decades systems biology has been pushed forward by the advent of high-throughput technologies for measurement of cellular components at large scale. Gene sequencing for whole organisms, facilitated that several model organism whole genomes were already available in the final years of the last century[69–72]. Later, next-generation sequencing technologies facilitated myriads of studies characterizing the RNA of cell samples[73,74], even at the single cell level[75,76].

Mass spectrometry methods have matured enough to be used for detection and quantification of the thousands of proteins inside cells[77–83]. Furthermore, this technology is applied to the detection of intracellular metabolites, nonetheless, this task remains challenging as these molecules tend to be small, present in low concentrations and have high turnover rates[11,84,85]. Additionally, carbon isotope labelling techniques have been used for following the breakdown of molecules by cellular metabolism to compute fluxes, using a model as a scaffold[86–89].

Altogether, these technologies have provided the means for the generation of a multitude of datasets of biological components measurement at a genome-scale. The study of these datasets is called -omics (genomics, transcriptomics, proteomics, metabolomics and fluxomics), a neologism indicating an entire set or a whole sphere of activity[90].

Throughout the development of systems biology as a field in biological sciences, two main approaches have been developed and implemented, the top-down and bottom-up (**figure 6**). The **top-down** approach consists of studying large datasets of experimental observations of molecular components (i.e., a bird's eye to the organism as a whole[91]), to extract information or construct explanations at higher levels of organization (pathways, cellular function, etc.). Based on an inductive process, in top-down systems biology methods, general or basic behaviors are sought through understanding or establishing connections of particular observations. Basically, this is a data-driven process, in which new biological information is extracted from large datasets[92].

Different quantitative techniques are available for the study of these datasets, for instance statistical characterization and modeling, such as differential expression analysis and gene-set enrichment analysis, for study of transcriptomics and proteomics data [71,93–95]; linear regression and correlation studies, for integrating multiple layers of data[96–99]; graph theory for metabolome studies[100–103]; and heuristic models such as neural networks, for study of transcription factor networks[104,105]. These approaches to science are often described as hypothesis-free or data-driven and are powerful tools for finding patterns in biological systems[106], nevertheless, it is hard to justify the generality of these findings, as for any inductive reasoning (known in philosophy of science as Hume's problem of induction[24]).

The other methodological approach in systems biology, **bottom-up**, consists of building mathematical models, usually based on basic principles or *a priori* fundamental knowledge, for description of biological systems and quantitative study of their properties and behavior[11,78]. Mathematical models enable hypothesis-driven science, by predicting the outcome of a system under perturbation in quantitative terms, allowing comparison with experimental information.

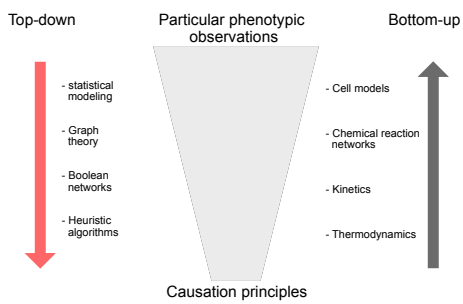


Figure 6.- The two methodological approaches in systems biology. The top-down approach represents an "eagles-eye view" over data, coming from particular observations, and aims to build knowledge on cellular mechanisms by finding relations among the data. Bottom-up systems biology relies on the construction of mathematical models, from basic general principles, that can explain particular cellular behavior, at different scales, in a quantitative fashion.

Broadly, modeling frameworks in systems biology can be classified according to the specificity of the mechanisms being abstracted. Thermodynamic and kinetic models describe interaction between cellular components, usually at the reaction or even quantum level, using fundamental physico-chemical principles. Chemical reaction networks are used to describe flow of matter and energy on larger scales (e.g., metabolic pathways, cellular compartments, or cellular subsystems). Finally, whole cell models attempt to provide a quantitative description of the cell as a whole, by integrating multiple biological phenomena of distinct nature, usually simplifying highly complex processes like protein synthesis.

1.6 A brief history of modeling in systems biology

Bottom-up systems biology has been formed by the contribution of ideas and methods coming from diverse fields, overall, with the intention of making biology a quantitative and predictive science. The pioneering ideas on systems thinking of Norbert Wiener[107] and Ludwig Von Bertalanffy[33,108] came in the 1950's; Denis Noble, a physiologist, developed the first model explaining the interaction between genetic components and properties at the cellular and tissue levels, the heart pacemaker model[31,32]; Mathematically-oriented chemical engineers, such as Arnold Fredrickson and Henry Tsuchiya, focused on explaining population dynamics of microbial cells in bioreactors during the 1960's and 1970's, using concepts like structured modeling, conservation equations, and differential equation systems with probabilistic terms[109–112].

The early 1970's also saw the rise of the metabolic control analysis theory and applications, mostly driven by biophysicists and biochemists like Reinhart Heinrich, Tom Rapoport and Henrik Kacser, during the 1970's[37,39,41], and Douglas Kell and Hans Westerhoff extending it in more recent times[113–116]. In the 1980's, another pivotal contribution from chemical engineering ideas came into systems biology, the formal analysis of reaction networks[117,118], applied by E. T. Papoutsakis[119] to transform metabolic pathway maps into algebraic systems, and by M. R. Watson, who developed a computational implementation of linear programming for obtaining numerical solutions of metabolic steady-states[120].

In the second half of the 80's Bernhard Palsson used dynamic models of reaction kinetics to understand metabolism, from single reactions[121] to a human red cell[122]. It was also Palsson's group, who in the early 1990's applied the concepts of steady-state reaction networks and optimization to the metabolism of the bacterium *Escherichia coli*, in a series of studies predicting its biosynthetic capabilities[123–126]. The first metabolic network at a genome-scale was published in 1999, for the bacterium *Haemophilus influenzae*[127]. Also during the final years of the XX century, new studies aimed to extend MCA and kinetic modeling to account for more complex reaction networks[128–130], non-linear kinetics[131,132], thermodynamic constraints[129,130], and the study of metabolic oscillations at the pathway level[13,133–137].

On parallel to these developments, other researchers were interested into the behavior of whole cells and pursued the project of modeling the interaction between different subcellular systems. Thus, a simple whole cell model for *E. coli* was published in 1979[138], and a more sophisticated one, including interactive software, in 1999[139]. This approach evolved further, to integrate more, and more sophisticated, description of subcellular processes, also implemented different mathematical formalisms for the process involved, this delivered the whole cell for *Mycoplasma genitalium*[140,141].

The different lines of thought that shaped bottom-up systems biology have consolidated into three main categories of quantitative modeling frameworks in systems biology: constraint-based methods, kinetic modeling, and multi-scale models. I follow this distinction as it allows for clear differentiation of models in terms of their biological scope and, or scale. Additionally, the three categories also display a clear distinction between the mathematical elements that define them, and the methods used for quantitative simulation. A summarized description of this is given in **table 2**.

Table 2.- Quantitative modeling frameworks in systems biology

Modeling framework	**Scope and scale**	**Mathematical formulation**
Constraint-based	Metabolism at a pathway, subsystem or genome-scale	Linear and non-linear equation systems
Kinetic	Metabolism or synthesis of macromolecules at the reaction, pathway or subsystem scale	Differential equation systems
Multi-scale models	Metabolism, synthesis of macromolecules, cell cycle, whole cell models	Mixed. Each cellular subprocess is modeled following a specific formalism and the different layers can be coupled.

Whole cell models proved to be powerful at reconciling multiple layers of omics data, while providing mechanistic connection, nonetheless, their formulation requires hundreds to thousands of molecular parameters, not readily available for a wide variety of organisms. Therefore, their applicability has been limited to three different species (*E. coli*[138,139], *M. genitalum*[140,141] and *S. cerevisiae*[142]) in 45 years of development.

Kinetic models provide detail predictions on the dynamics of chemical systems. Their detailed representation of biochemical reactions, based on kinetic mechanisms, enables construction of equation systems that predict metabolite concentrations and reaction fluxes. However, the complex kinetic expressions require the incorporation of extensive enzyme parameters. Additionally, the non-linearity and the common coupling in between mass balances across metabolites, demands the use of complex

mathematical methods or approximations for obtention of quantitative prediction, hindering their application to large-scale metabolic networks, usually restraining their use to the pathway and reaction level.

1.7 Genome-scale metabolic models and flux balance analysis

A genome-scale metabolic model (GEM) is a computational representation of an organism's metabolism, encompassing all the biochemical reactions that occur within its cells. It serves as a comprehensive blueprint of the chemical transformations that enable the organism to grow, produce energy, and carry out various functions. GEMs are constructed using information from the organism's genome, cataloging the enzymes and transporters encoded in its DNA and mapping them to the reactions they catalyze[143]. They represent a knowledgebase on cell's metabolism, offering a comprehensive catalogue of its molecular components and their interactions.

One powerful technique employed with GEMs is Flux Balance Analysis (FBA). FBA is a computational method used to simulate and analyze the flow of metabolites through the metabolic network of an organism under steady-state assumption, where the rates of metabolite production and consumption are balanced (i.e., no internal accumulation of metabolites mass)[144]. This assumption enables the construction of mass balances around each metabolite, mathematically represented as a homogeneous system of linear equations. Nevertheless, as metabolism is highly interconnected and redundant, these systems tend to be highly underdetermined (i.e., non-zero degrees of freedom). Due to this FBA operates on the principle of optimizing a cellular objective while accounting for mass conservation and thermodynamic constraints, as undetermined systems are characterized by a solution space of infinite solutions, rather than a unique one. A commonly used objective function in FBA is to assume that cells allocate resources to maximize their growth rate[145].

In practice, FBA uses linear programming to find the distribution of reaction fluxes (the rates at which reactions occur) that satisfy the constraints of the metabolic network to achieve the imposed objective, thus FBA is at the core of constrain-based methods in systems biology[146]. By solving this mathematical problem, FBA predicts how nutrients are utilized, how byproducts are generated, and how metabolic pathways are coordinated to support cellular function[144].

Researchers utilize GEMs and FBA in various ways, ranging from understanding an organism's metabolic capabilities to guiding the design of biotechnological processes. In the context of biotechnology, GEMs can aid in the rational engineering of microorganisms for the production of valuable compounds such as biofuels, pharmaceuticals, and chemicals[147]. By manipulating the reaction fluxes in silico, researchers can predict and fine-tune metabolic engineering strategies to optimize product yields and reduce unwanted byproducts.

1.8 Protein-constrained models

The determination of accurate flux distributions using GEMs and FBA is a major challenge, as optimal values of an objective function can be attained by alternate flux distribution profiles[148]. This means that a global systems behavior, such as the growth rate of a cell, can be realized by different intracellular metabolic states, however, not all the alternate optima in an FBA problem are biologically meaningful. Therefore, the quality number of constraints defines the accuracy of predicted phenotypes. Nonetheless, intra and extracellular flux constraints are not readily available for a wide variety of organisms and conditions.

The concept of cellular resource allocation has been explored to incorporate additional constraints on metabolic models. This framework mainly consists of considering a resource known to be limited in the cell, and then formulate a resource cost for each reaction of the network. Consequentially, phenotypes get constrained by flux of nutrients, stoichiometry, and a finite cellular resource. These approach has been used by considering a crowded cellular environment[149,150], a finite membrane area for the expression of transporter proteins[151], and bounded total protein mass available for metabolic enzymes[152–158]. These modeling frameworks have succeeded at refining the phenotype predictions of classical FBA for growth in diverse environments. Furthermore, this kind of models are capable of capturing overflow metabolism in *E. coli*, *S. cerevisiae* and human cells.

These models connect reactions in the network to the constrained resource by establishing a flux cost. In models constrained by the total amount of protein such cost becomes the enzyme demands of each reaction. The use of the Michaelis-Menten equation enables representation of reaction fluxes as a function of enzyme amounts:

$$v = V_{max}\frac{S}{S+K_m} = k_{cat}[E_0]\frac{S}{S+K_m} \qquad \text{(eq 1.1)}$$

Where V_{max} is the maximum attainable flux by a reaction; S is the concentration of the substrate; K_m is the Michaelis constant; k_{cat} is the turnover number of the enzyme; and $[E_0]$ is the concentration of the enzyme. Consequently, this expression imposes an upper limit on reaction fluxes, which cannot be higher than the product of k_{cat} and the concentration of the enzyme.

Aims and significance

In the process of writing this thesis, and reviewing the research done during my PhD studies, I have identified two central elements among the included publications: a habit of looking at biological phenomena from different angles and perspectives to build understanding, and constant effort to leverage systems biology methods to deliver both applied solutions and conceptual progress, sometimes opening up new questions.

In **paper I**, the phenotypic responses in budding yeasts, caused by environmental changes are investigated at a global molecular level. Bioinformatics methods enabled identification of stress responsive genes as evolutionarily young genes. Results were summarized in a proposed hypothesis for evolution of long-term stress adaptation mechanisms in budding yeasts. This study shows an example of the iteration between proximal (e.g., phenotypic changes due environmental factors) and distal causation (biological causation caused by changes in the genotype across reproduction events) in biological systems. Demonstrating that observable phenotypes are the result of diverse phenomena (environmental, physiological, and evolutionary) interacting at multiple scales and time spans.

Paper II presents a critical assessment of the use of GEMs for studying diverse yeast species. A catalogue of all the available GEMs for budding yeasts was collected and made available for the community. Furthermore, a basic testing pipeline was also developed and used to evaluate GEMs. Results showed how accessibility, usability, and interoperability of GEMs can be benefited by incorporating modeling standards, version-control and community development practices. These concepts were central to the development of yeast 8[159], the consensus model of *S. cerevisiae*, and humanGEM[160], the most comprehensive model of global human metabolism available, projects in which I also contributed to during the time of my PhD studies, but remained outside the scope of this thesis.

Paper III provides an extension to the framework of enzyme-constrained metabolic modeling. By updating GECKO to its 2.0 version several objectives were reached, namely, generalization of the software to GEMs for any organism; circumvent the problem of the lack of kinetic data for specific organisms; provide an automatically updated catalogue of high-quality ecGEMs for five different organisms, facilitated by software version-control and continuous integration.

Additionally, a case study on integrative data analysis is presented, in which proteomics data, generated for **paper I,** were integrated into ecGEMs for *S. cerevisiae, Y. lipolytica* and *K. marxianus*. Results indicated that these yeasts have evolved different molecular mechanisms to tolerate environmental stress, despite their phylogenetic relation. Nevertheless, a possible systems-level emergent property was identified, up-regulation and high saturation of enzymes in different sectors of amino acids metabolism as a common stress-response mechanisms across budding yeasts.

Systems biology applications for development of technology are introduced in **papers IV** and **V**. In the former, enzyme-constraints are used for improving predictions of metabolic engineering targets by traditional GEMs and methods. Gene targets for increasing production of heme by *S. cerevisiae* cells were predicted, the top gene modification predictions were incorporated into a yeast strain, yielding a 70-fold increase in the intracellular heme levels. Here, the possibility of improving model predictions by incorporation of kinetic data, was achieved and results delivered a promising production strain for a highly demanded product.

The method for prediction of gene targets for increased bioproduction was further improved and systematized, focusing on providing optimal combinations of gene targets and facilitate understanding of the predicted engineering strategy. In **paper V** production of 102 diverse chemicals in *S. cerevisiae* was simulated with an ecGEM; lists of predicted targets for each product were made available for web-based visualization. Moreover, sets of targets predicted for groups of products, instead of individual chemicals, were identified, suggesting the possibility of model-driven development of platform strains. Finally, general principles for guiding metabolic engineering in yeast cells were extracted from analyses over simulations and predicted targets.

These two applied studies demonstrate the value of using mathematical models for guiding bioengineering projects, as they provide an initial tool for prediction of intervention strategies, but also a conceptual scaffold for analysis and understanding. Altogether, activating the design-build-test-learn cycle of synthetic biology and metabolic engineering.

In **paper VI** integration of multiple modeling formalisms is used to represent the crosstalk between nutrient signaling, gene expression and metabolism in *S. cerevisiae*. The hybrid model provided explanation to cellular phenotypes that cannot be predicted by any of the implemented modeling tools individually. Namely, enzyme expression profiles that do not correspond completely to catalytic or substrate optimization. Predictions explain how different layers of cellular processes interact and give rise to multi-purpose cellular phenotypes. As suggested by the interplay between upwards and downwards causation in systems biology.

In summary, this doctoral thesis presents a series of studies aiming to understand the complexity of cellular phenotypes in budding yeasts, using the methodological flexibility of systems biology as a means. The fluid nature of systems biology research, oscillating between basic and applied science, set a framework for delivery of technological and conceptual outputs in this work. Scientific and technological assets, in the form of model catalogues, software tools, and methods for engineering, are presented and made publicly available. On parallel, new hypotheses, mechanistic explanations, and even new questions emerged from this body of work.

2.- Understanding long-term adaptation to environmental stress in budding yeasts: A top-down approach

The budding yeasts *S. cerevisiae, K. marxianus* and *Y. lipolytica* have acquired a variety of niche-specific adaptations that make them interesting for bioproduction in industrial conditions. *S. cerevisiae* displays high tolerance to ethanol and osmotic pressure[161]; *K. marxianus* is able to grow at high rates in high temperatures[162]; both of these conditions can be found in industrial processes. Additionally, *Y. lipolytica* has evolved to endure hydrophobic environments, and is also known for its high accumulation of lipids, also interesting for bioproduction[163].

It is desired to engineer these yeast strains to transform them into robust cell factories for diverse production of chemicals, as an alternative to conventional chemical processes. In this context, robustness implies providing the yeasts with production and proliferation capabilities up to the industrial level. For this, it is crucial to understand the behavior of the different yeast species under different environments, provided with particular stress factors[164]. In this study low pH, high osmotic pressure (characteristic of raw material feedstocks), and high temperature (common in industrial processes) are probed as stress conditions, to answer the following questions:

1.- Are there any multi-stress responsive biological mechanisms in any of these yeasts? (same mechanism in several conditions, for a given yeast)

2.- Are there any biological mechanisms for stress tolerance shared by these yeasts in particular conditions? (same mechanism across yeasts, for a given condition).

2.1 Experimental characterization of stress responses

In order to provide a comparable set of conditions to test for these questions, growth rate was kept constant in controlled chemostats, where cell cultures were grown at 0.1 h^{-1}, biological triplicates were ran for each organism-condition pair. Conditions are specified in **table 3**.

Table 3.- Experimental conditions for chemostat cultures at0.1 h^{-1}

Condition	***S. cerevisiae***	***K. marxianus***	***Y. lipolytica***
Standard	30 °C, pH 5.5	30 °C, pH 5.5	28 °C, pH 5.5
High temperature	36 °C, pH 5.5	40 °C, pH 5.5	32 °C, pH 5.5
Low pH	30 °C, pH 3.5	30 °C, pH 3.5	28 °C, pH 3.5
Osmotic stress	30 °C, pH 5.5, 600 mM KCl	30 °C, pH 5.5, 600 mM KCl	NA

The use of chemostats allows cell cultures to adapt to stress, through several generations, until the steady-state is reached. Thus, samples represent a long-term

adaptation phenotype, to not be confused with stress shock experiments caused by exposure to pulse variations in environmental variables[165].

RNA and total protein were isolated from cell pellets obtained from the samples for high-throughput quantification. RNA sequencing was performed using next generation illumina sequencing and returned transcript data in transcripts per million (TPM). Protein measurements were performed using LC-MS/MS and quantified using XIC (eXtracted Ion Current)[77,166]. Transcriptome and protein data were processed in similar ways, undergoing steps of removing extreme low counts and non-consistently measured elements, TMM normalization for transcriptomics, and sample quality assessment through PCA. This data treatment was necessary to perform differential expression analysis of transcripts and proteins in a fair way, addressing how much the abundance of a given gene product varies in stress conditions in comparison to a control condition.

2.2 Stress responsive mechanisms across conditions

Statistically significant genes, at the transcript level, were identified using $log_2FC = \pm 2$ and $FDR < 0.01$ as significance threshold values. **Figure 11** shows that for the three organisms, a big proportion of the stress responsive genes are condition specific, and few were found to be responsive to multiple conditions, in comparison. Proteomics measurements were also assessed for differential expression, showing similar patterns, however with a much lower number of DE proteins in comparison to transcripts, even when relaxing FDR upper limit to 0.05.

In order to understand possible system mechanisms involving the differential expressed genes, gene function was sought for all organisms, querying from Ensembl database[167] for the *S. cerevisiae* case, and BLAST2GO[168] for the other yeasts, for which each protein-coding genes searches for a functionally annotated homolog. This process failed to annotate 20% of the *K. marxianus* measured mRNAs, 38% in the *Y. lipolytica* case. In the case of *S. cerevisiae*, around 11% of measured transcripts lack a functional annotation (**figure 12A**).

Mapping the DE genes, at the mRNA level, to the lists of functionally annotated genes revealed that unannotated genes were overrepresented in all conditions for the three yeasts (**figure 12B**). This proportion reached 50% in *Y. lipolytica*. These results suggested that performing gene-set enrichment analysis among the DE genes, in order to characterize stress adaptation mechanisms, would not be representative of the whole transcriptional response of the yeasts. Additionally, proteomics measurements failed to quantify a big proportion of proteins encoded by genes that were DE at the transcript level, suggesting that they encode for very low weight proteins, usually elusive to MS/MS approaches[77], or were discarded by the previous filtering steps of data processing, due to inconsistent measurements across biological triplicates in a given condition.

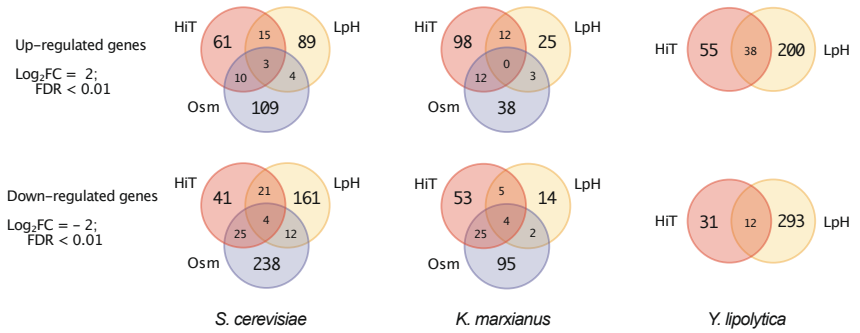


Figure 11.- Statistically significant differentially expressed genes in *S. cerevisiae*, *K. marxianus* and *Y. lipolytica*, after long-term exposure to high temperature (HiT), low pH (LpH) and osmotic pressure stress (Osm).

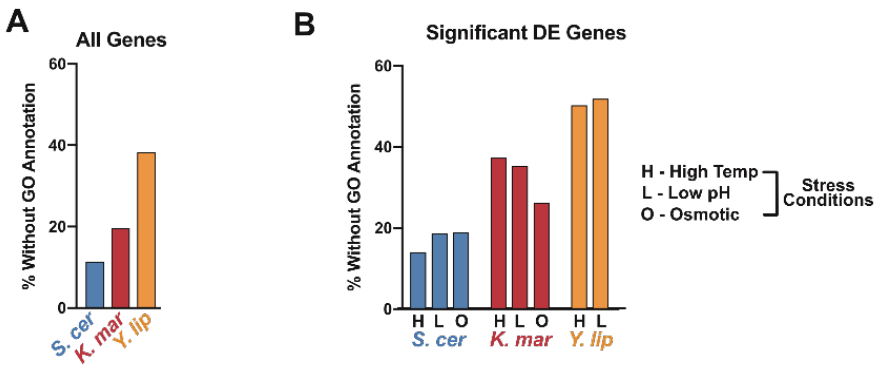


Figure 12.- Genes without functional annotation in *S. cerevisiae*, *K. marxianus* and *Y. lipolytica*. A) Proportion of unannotated genes among all genes. B) Proportion of unannotated genes among the total number of differentially expressed genes (up-regulated + down-regulated) by stress condition.

2.3 Are transcriptional stress responses evolutionarily conserved across budding yeasts?

Then, for addressing the second of the initial questions, the core genes across the three yeasts, thought to be present in the least common ancestor between them, around 325 MYA[169] (**figure 13A**), were identified running homology searches of single copy orthologs among their amino acid protein sequences, using orthoFinder software[170]. A list of 2959 core protein-coding genes was obtained, and DE genes for the conditions shared by the three yeasts (high temperature and low pH) were mapped to it. A low fraction of the total DE genes found a match in the list of single-copy orthologs. Strikingly, almost none of the identified matches showed to be DE in more than one yeast species, shown in **figure 13B**. This indicates that the response

to long-term high temperature and low pH exposure is mostly by differential expression of non-core genes, or genes that emerged later than their LCA.

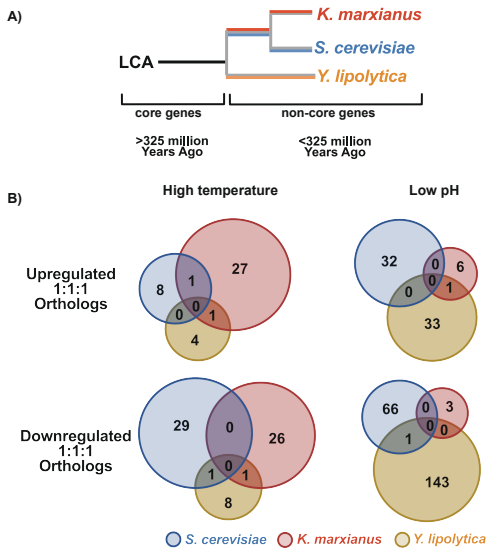


Figure 13.- Core genes shared by *S. cerevisiae, K. marxianus* and *Y. lipolytica*. A) Simplified phylogenetic tree indicating the speciation event separating the clades of the three yeasts. Single-copy orthologs across the three species are thought to have been present in their LCA. B) Mapping of DE genes in high temperature and low pH to the set of core genes. Venn diagrams show DE of genes in a given species, condition and directionality. Stress responsive core genes are mostly restricted to a single species.

A least stringent approach was used to classify non-core genes, aiming to account for duplication events of genes. Therefore, genes present as multiple copies in a given yeast, and as one copy in the other two species, were classified as multi-core genes. In this way all the protein coding genes of the three yeasts were divided in three classes: single-core genes, multi-core genes, non-core genes, as shown by **figure 14A**. As an example, HIS1 is a single-core gene; GAL1 and GAL3 provide a case of a duplicated gene in *S. cerevisiae* present as a single copy in the *K. marxianus* and *Y. lipolytica* (GAL1); while the gene YJL199C is only present in *S. cerevisiae* (**figure 14B**). In general, single-core genes account for most genes, detected at the transcript level, in the three yeasts, more than 50% in all cases, showing even higher values among the protein level measurements (**figure 14C**).

Mapping of the DE expressed genes, from all organism-condition pairs, and both at the transcript and protein level, revealed a clear pattern of a significant highly likelihood of DE, caused by stress exposure, of younger and duplicated genes when compared to evolutionarily conserved ones, or "older genes", as shown in **figure 15**.

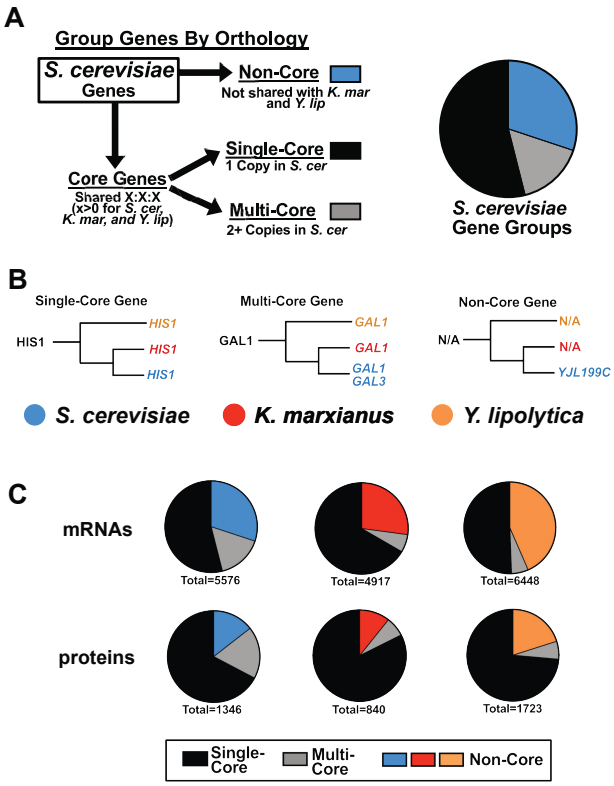


Figure 14.- Gene grouping according to conservation among the three yeasts and duplication events. A) An example of the gene grouping rationale, explained for the *S. cerevisiae* case. B) Examples of single-core (HIS1), multi-core (GAL1-GAL3) and non-core genes in *S. cerevisiae* (YJL199C). C) The total number of protein-coding genes, sorted by conservation and duplication categories, detected at the mRNA and protein level for *S. cerevisiae, K. marxianus* and *Y. lipolytica*.

As the evolution of the Saccharomycotina subphylum spans 400 million years of evolution, and these 3 species span a large swath of its diversity, a deeper look into the timing of gene emergence/duplication and their likelihood to be DE was necessary. To assess *de novo* gene emergence, for each species, each protein-coding sequence of AA was searched for homologs across 3 other species at the genus, clade, subphylum and phylum levels using orthoFinder (**figure 16A**). A similar process searched for duplication events in the evolutionary history of each species (**figure 16B**). **Figure 16C** provides an example of the phylogeny of the proteomes that were used for homology searches with orthoFinder for the *S. cerevisiae* case. Details on this sorting algorithm are further explained in the **supplementary material** of **paper I**.

With this approach the protein-coding genes of the three yeasts were divided into 6 groups, genes conserved at the phylum level (group I), subphylum (group II), clade (group III), genes (group IV), and genes restricted to the species-strain level (group V) (**figure 16D**). For *S. cerevisiae* an additional group of genes was introduced in between the clade and genes level, corresponding to a major duplication event (called the whole-genome duplication event, WGD, in the literature) occurring 25 MYA. The increasing number in the group names represents how evolutionary "young" genes, or gene duplication events, are.

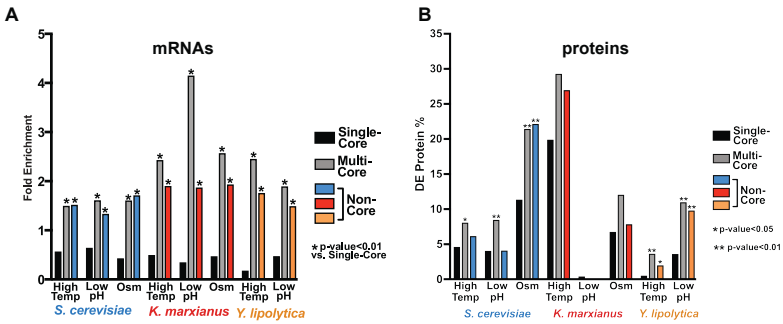


Figure 15.- Multi and non-core genes are enriched for stress-responsive DE genes. A) Normalized ratio of the total number of genes per group, over the number of DE genes, at the mRNA level, in each group. Results were normalized by the ratio between all DE genes over all genes quantified for each organism-condition. *p-values*, under a test, indicate significant differences between the enrichment found multi and non-core genes vs. core genes. B) Percentage of DE proteins out of the total number of detected proteins per group, shown for the three yeasts and all stress conditions. *p-values*, under a test a two-sided Fisher's exact test, indicate significant differences between the percentages for multi and non-core genes vs. core genes.

The refined gene grouping confirmed that stress responsive genes (at transcript level) are significantly underrepresented, proportionally, among the evolutionary conserved genes. Therefore, young genes are enriched for stress-responsive genes in these conditions and species, specifically those corresponding to groups IV, V and WGD in *cerevisiae*, the genus and species-specific genes and gene duplication events. This enrichment is as high as four to six times higher for group V genes, in comparison with what would be expected by chance (total number of DE genes over the total number of genes). These results are shown for the *S. cerevisiae* case in **figure 17A**, and a very similar pattern was observed for the *K. marxianus* and *Y. lipolytica* cases, as shown by figure 3 and supplementary figure 7 in **paper I**, respectively.

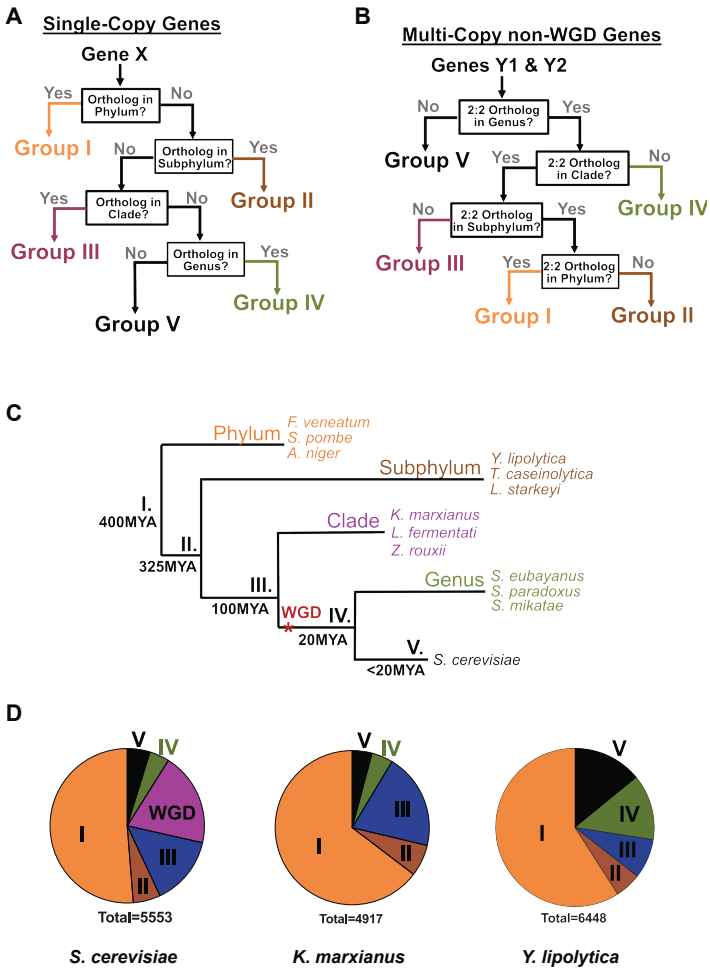


Figure 16.- Gene sorting according to evolutionary age and duplication events. A) Bottom-up approach followed for assessing evolutionary age of single copy genes. Starting from phylum, the first match as of a gene as ortholog with the organisms in a given phylogenetic level assigns its belonging to an age group. B) Top-down approach followed for multi-copy genes. Starting from genus, and going downwards, a duplication event is established as the last or lower phylogenetic level in which orthologs with the same number of copies can be found for each gene. C) Simplified phylogenetic tree used for sorting of the *S. cerevisiae* genes, each phylogenetic level shows the organisms that were selected as representative for it. The same kind of trees were reconstructed for *K. marxianus* and *Y. lipolytica*. D) Total number of genes measured at the transcript level sorted by emergence and duplication timing for the three yeast species.

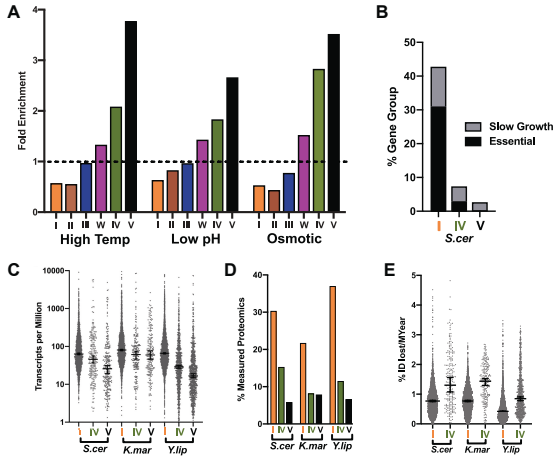


Figure 17.- Comparison of genes and their proteins across age groups. A) Fold enrichment for DE genes across gene groups for *S. cerevisiae* in all stress conditions. B) Percentage of essential and growth-related genes in groups I, IV and V for *S. cerevisiae*. C) Distribution of gene expression levels (at transcript level) in standard conditions across gene groups I, IV and V for the three yeast species. D) Distribution of the percentage of AA sequence identity loss per million years for the genes in groups I and IV in the three yeast species.

2.4 Young genes are lowly expressed and encode for non-essential and rapidly mutating proteins

Additional experimentally characterized information was searched in the literature for the genes of *S. cerevisiae*. Investigation of cellular localization (according to localization GO terms[53]) of the proteins encoded by young genes (groups IV and V) revealed a significant enrichment for localization in the plasma membrane, cell wall and vacuole. In contrast, proteins encoded by genes in group I were significantly enriched (Benjamini-Hochberg corrected p-values[171], computed by a hypergeometric test) for mitochondrial, cytoplasmic and nuclear localization. More than 40% of ancient genes were found to encode for essential proteins, or proteins impairing growth under deletion, whilst this proportion is very low in comparison with genes in groups IV, V (**figure 17B**). Together, this suggests the involvement of ancient genes in core cellular function, and young genes being related to less essential processes, most likely involving extra/intra cellular and intercompartmental exchange.

These results were complemented by the observation that, on average, genes in groups IV and V are expressed as mRNA in lower levels than those in group I, across the three yeasts growing on standard or reference conditions (**figure 17C**). Also, the percentage of genes detected at the protein level in non-stress samples by MS/MS, decreases drastically for young genes, in comparison to ancient ones (**figure 17D**). Failure of protein detection in MS/MS approaches has been reported to be related to proteins with very low abundance in the measured samples[166]. Additionally, the data

processing steps may have discarded proteins with highly variable abundances across biological triplicates.

Other studies have found a relation between low expression levels and non-essentiality of gene products with increased mutation rates in *S. cerevisiae*[172] and *Schizosaccharomyce pombe*[173]. This motivated the analysis of the mutation rate of genes for the yeasts in this study. For each yeast species, AA sequence identity was compared between homologous proteins from members of the same genus, adjusted to the estimated evolutionary time elapsed between each pair of species (16.9 MY between *S. cerevisiae* and *S. eubayanus*; 27.43 MY between *K. marxianus* and *K. lactis*; and 22.42 MY between *Y. lipolytica* and *Y. bubula*[169]), allowing to assess for adaptation rates of protein sequences. This showed that, on average, younger genes, shared at the genus level, display significantly higher adaptation rates than ancient ones (group I).

Based on all the finding in this work, we propose a model of evolution to intermittent stress, in which random mutations may occur among all genes, then those mutants that cannot grow properly will be eventually selected out (counter-selection of growth deficient mutants). From the remaining pool of mutants, those that are unfit for enduring the environmental factors are again counter-selected (non-beneficial mutants). At the end of this cycle, the stress-tolerance beneficial mutants remain in the population and, after generations, can pass these mutations to new cells.

2.5 Summary

In this chapter, the approaches followed to study long-term adaptation to stress by *S. cerevisiae, Y. lipolytica* and *K. marxianus*, are described. Transcriptomics and proteomics data were obtained from bioreactor cultures at steady state, using a low dilution rate. Differential analysis of the transcriptome and proteome revealed that stress-responsive genes tend to be uncharacterized genes, specially for *K. marxianus* and *Y. lipolytica*. Moreover, it was found that molecular tress responses of budding yeasts are niche-specific, and not majorly shared across the Saccharomycotina subphylum.

Confronted with these results, an approach for dividing the entire genome of these yeasts into gene groups that reflect the evolutionary or duplication age, was developed. Mapping of DE genes, at the transcript level, to the obtained gene groups, showed that young genes (restricted to the genus or species) are more likely to be DE under stress, in comparison with ancient conserved genes. Further analysis showed also that young genes are lowly expressed as mRNA and tend to encode for proteins that 1) are not involved in cellular growth or essential processes; 2) are elusive to MS/MS detection, suggesting heterogenous or low expression; 3) are located in cell wall and plasma membrane; and 4) display higher rates of mutation than ancient genes.

All these results enabled the proposal of an evolutionary mechanism for long-term stress response, seemingly conserved across budding yeasts. Notably, the resulting hypothesis was purely derived from the data analysis, without the need of introducing any directionality or teleology in the evolutionary hypothesis. This shows an example of the possibility of gaining understanding of high-level cellular or even evolutionary processes from an Eagle-eye look into high-throughput molecular data, even in the absence of extensive annotation for gene functionality.

From a more practical point of view, the findings in **paper I** provide the metabolic engineering and synthetic biology communities with a list of uncharacterized genes correlated with stress adaptation and young evolutionary age, and seemingly not related to essential processes. Past studies have found that engineering gene expression for stress-tolerance purposes usually comes with a trade-off of other cellular functions. Thus, the stress responsive young genes identified here constitute a set of candidate targets for improving strain robustness.

3.- Genome-scale metabolic modeling of budding yeasts: evaluation of accessibility, usability and interoperability

The yeast *S. cerevisiae* was the first ever Eukaryote genome to be sequenced in its entirety in 1996[70]. This, together with the accumulated molecular biology data for this organism in public databases, made possible to reconstruct one of the very first genome-scale metabolic models, iFF708, in 2003[174]. This model was utilized for comparison of biosynthetic capabilities against the *E. coli* network.

For model reconstruction for other less studied species, metabolic modelers developed a model reconstruction approach, in which a well-curated preexisting network, for a phylogenetically related organism, is taken as an initial model scaffold. The draft model is then refined with computational algorithms, addition of molecular data and homology searches for genes encoding for functionally annotated enzymes. This approach proved to be fruitful for modeling the metabolism of other non-cerevisiae budding yeasts, as the iFF708 model, and its progressive improved versions, provided a comprehensive draft model to guide this process.

The history and genealogy of almost 20 years of yeast models has been extensively reviewed by other authors[175–178], therefore, in **paper II**, an effort for collecting, comparing and evaluating the high-quality published models of diverse yeast species was done. In total, 45 different GEMs for 12 different yeast species were found in the literature (shown in **table 4**). Model files were sought in the web and collected into a single repository, publicly available at: https://github.com/SysBioChalmers/YeastsModels.

Table 4.- Number of high-quality GEMs reconstructed per yeast species (2003-2021).

Name	**# of models**
Saccharomyces cerevisiae	19
Komagataella pastoris	8
Yarrowia lipolytica	5
Scheffersomyces stipitis	4
Rhodotorula toruloides	2
Candida glabrata	1
Candida tropicalis	1
Kluyveromyces lactis	1
Kluyveromyces marxianus	1
Lachancea kluyveri	1
Schizosaccharomyces pombe	1
Zygosaccharomyces parabailii	1

As expected, *S. cerevisiae* is the species with the highest number of models available (18). It was also observed that different reconstructions for a given species are

usually developed by different research groups. Even though a series of so-called consensus models have been sequentially published for *S. cerevisiae*[159,179–183], new reconstructions stemming from independent efforts have also been released. Different groups usually have differences in their scope and questions, however, this finding may also indicate that researchers do not always choose to leverage the knowledge collected by previous attempts, which very often results in model inconsistencies.

It was also found that the oldest models for *S. cerevisiae* are still being highly cited, in some cases even more than more recent publications or even the consensus models, shown in **figure 18A,** and figure 2F and the supplementary materials of **paper II**. If citations can be seen as a proxy measure of model utilization, this suggests that the community is used to exploit previous model versions, instead of leveraging the systematization and correction of knowledge available in more recent models.

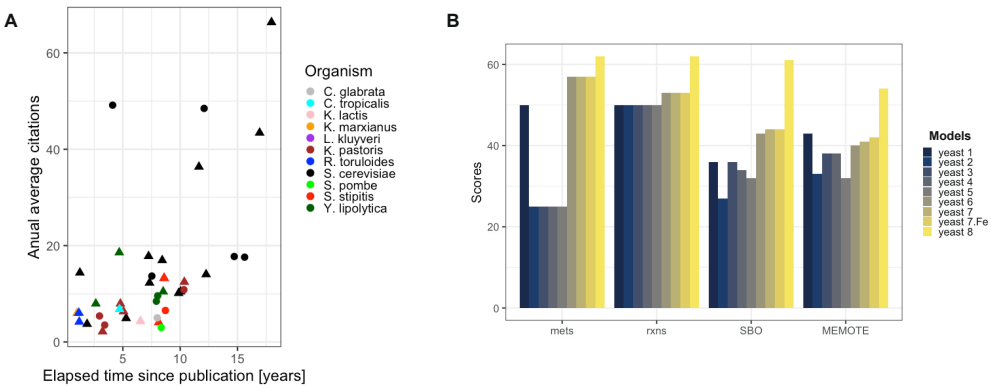


Figure 18.- Usability of budding yeasts GEMs. A) Citation landscape of GEMs for diverse yeast species. B) Evolution of the Memote scores for annotation on different model fields across versions of the consensus yeast model.

The process of model recollection revealed that models are made public in different ways and formats. Most of the models were accessible as part of the supplementary material of their corresponding publications, and 55% of them were deposited in specialized databases for metabolic models, such as biomodels[184] and openCOBRA[185]. Nonetheless, these models are available in a wide variety of file formats, despite the existence of a standardized format for model reconstruction and sharing, the Systems Biology Markup Language (SBML)[186]. For 26% of the models no SBML file was found in the web and, instead, these models were shared in alternative software-dependent formats which may not be compatible with most modeling simulation software available.

Figure 18A also shows that in the last 10 years, more yeast species have been modeled using GEMs, however, multiple reconstructions exist for other species that are of relevant industry for research and industrial purposes. Which raised concerns about the confusion that researchers may face when needing to select a model for their applications. Inspired by this, we developed a simple computational pipeline that simulates the first use of a GEM by an unexperienced user. This pipeline has the objective of identifying the most relevant model components to run a simulation and attempts to solve for a simple FBA problem using the COBRA[187], RAVEN[188] and COBRApy toolboxes[185].

43 models were tested with this pipeline (as model files could not be found in any source for two of the published models). Most of the models were found to be readable by either of the mentioned toolboxes. For 24% of the studied models no preestablished objective function was found. Similarly, a biomass or growth pseudoreaction could not be found easily within the model structure for 16% of the cases (searching for preestablished objective or querying the most used names for this reaction in the literature). Additionally, it was not possible to run a successful cellular growth simulation, using the preestablished constraints and available information, for 24% of these models. Therefore, the process of familiarization with a specific model structure is not straightforward for a considerable number of these models.

Furthermore, when running analysis with GEMs it is a common task to search for evidence of reactions, metabolites and genes in other databases or other models. Therefore, annotation of model components with unambiguous identifiers and connections to external databases is essential for model utilization. The MEMOTE test suite[189] was used for evaluating the degree of annotation of model components for these 43 models, however, the test failed for 36% of them, due to inconsistencies or errors in the model file. Nevertheless, the MEMOTE test was successfully ran for all the models belonging to the *S. cerevisiae* series, enabling comparison across model versions. Results of this test, showed a consistent improvement in the annotation of metabolites, reactions, SBO terms (system biology ontology terms[190]), and consequently on the overall memote score of the consensus cerevisiae models across versions, with the highest scores obtained by the latest reconstruction yeast 8[159] (**figure 18B**). Suggesting that coordinated modeling projects, community development, and version-control practices (as in the case of yeast 8) are useful for improving model quality. A detailed summary of the test results (customized test + Memote) can be found as supplementary materials of **paper II**.

3.1 Summary

In **paper II** a critical assessment of GEMs for diverse yeast species was carried out. 45 models for 12 species were found among the scientific literature. Model files were

collected, offering a catalogue of GEMs for diverse yeasts. This process revealed that model files may not be readily accessible in the web, and that there exists variability regarding the file format that researchers choose for their models. This introduces problems to the modeling community, as non-compliance to standard formats creates compatibility problems with simulation and testing software for metabolic modeling. Additionally, it was also observed that redundancy in modeling efforts is common, as several model reconstructions have been published by different research groups for some of these species.

A simplified testing pipeline, aiming to simulate the use of a GEM by a new user, revealed that for 24% of the models, executing simulations provided with minimum information, was not possible. Additional tests showed how model quality, in terms of annotation and consistency, can get benefited from long-term, community-driven, cumulative modeling projects, as in the case of the series of *S. cerevisiae* consensus models.

The findings in **paper II** call metabolic modelers to comply to standard practices and formats for sharing models (such as maintained databases or git repositories); leverage community efforts and build upon previous modeling projects to maximize the amount of knowledge contained in GEMs; improve model accessibility (standard identifiers and annotation) to bring GEMs closer to a wider number of researchers and industry professionals.

4.- Extending the concept of enzyme-constrained metabolic modeling to multiple organisms

As mentioned in the first chapter, the development of efficient cell factories is a resource and time intensive process, which may take years and millions of USD to be completed. This process can be alleviated by reconstruction of GEMs, which provide an extensive knowledgebase of cell metabolism, a tool for quantitative simulation and prediction, and a scaffold for omics data integration.

There is high potential for developing efficient cell factories using the three yeast species studied in **paper I** as platforms, as they display interesting phenotypic adaptations to environments that can also be found in industrial settings. Additionally, there is considerable accumulated knowledge on engineering the metabolism of these yeasts for directed purposes, especially for *S. cerevisiae*.

From paper II, three candidate GEMs for these species were identified as suitable for guiding engineering attempts. The models yeast8 for *S. cerevisiae,* iSM996 for *K. marxianus*, and iYali4 for *Y. lipolytica*, were all found to be the products of community-development projects, including the feedback of supporting communities of users, reconstructed using version-controlled approaches, and comply to standard practices in model format (model format, annotation of model fields with external database identifiers).

Previously, incorporation of enzyme constraints proved to significantly improve the content scope and prediction capabilities of the model yeast7. In order to update the ecGEM of *S. cerevisiae* to its version 8, and reconstructing ecGEMs for *K. marxianus* and *Y. lipolytica*, it was necessary to update the GECKO toolbox, which was originally designed to work with the model yeast 7, and also suited for the high availability of kinetic parameters for *S. cerevisiae*.

4.1 The GECKO formalism for incorporation of enzyme constraints into metabolic networks

Enzyme-constrained produced by GECKO are members of the family of protein-constrained models, explained in **section 1.8**. GECKO accounts for enzyme cost of reactions by integrating enzymes for each reaction as if they were components in the model, they become a pseudometabolite in the model (*E*). The expression for the maximum flux attainable through a reaction is given by[16]:

$$v_i^{MAX} = k_{cat_{ij}}\left[E_j\right] \text{ (eq. 4.1)}$$

Where $k_{cat_{ij}}$ is the turnover number of enzyme *i* for the substrate in reaction *j*, in units of s^{-1}; and $\left[E_j\right]$ corresponds to the abundance of the enzyme in units of *mmol/gDw*. Therefore, the real demand of the enzyme *j* by reaction *i,* is expressed by:

$$e_{ij} = \frac{v_i}{k_{cat_{ij}}}, s.t. e_j \leq \left[E_j\right] \text{ (eq. 4.2)}$$

For each enzyme j, a total demand (e_j) is expressed as the sum of enzyme demands across all reactions in which it participates. By accounting for e_j as pseudoreactions in the model, mass balances can be constructed around each enzyme (E_j), which in steady-state take the form of:

$$\frac{dE_j}{dt} = e_j - \sum_i \frac{v_i}{k_{cat_{ij}}} = 0, s.t. e_j \leq \left[E_j\right] \text{ (eq. 4.3)}$$

This enables straightforward integration of the enzyme mass balances into the matrix formulation of a GEM, adding enzymes as new rows and enzyme usages as new columns in the stoichiometric matrix. The modified stoichiometric matrix then takes the following form for a reaction network with m metabolites, n reactions, and p enzymes:

$$S = \begin{bmatrix} S_{1,1} & \cdots & S_{1,n} & 0 & \cdots & 0 \\ \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ S_{m,1} & \cdots & S_{m,n} & 0 & \cdots & 0 \\ \frac{-1}{k_{cat_{1,1}}} & \cdots & \frac{-1}{k_{cat_{1,n}}} & 1 & \cdots & 0 \\ \vdots & \ddots & \vdots & \vdots & 1 & \vdots \\ \frac{-1}{k_{cat_{p,1}}} & \cdots & \frac{-1}{k_{cat_{p,n}}} & 0 & \cdots & 1 \end{bmatrix} \text{ (eq. 4.4)}$$

Thus, in steady-state, the following system of linear equations is obtained:

$$Sv = 0, \;\; v = \left\{v_1, \ldots, v_n, e_1, \ldots, e_p\right\} \text{ (eq. 4.5)}$$

In order to obtain flux distributions using FBA, constraints are set in the form of:

$$LB_i \leq v_i \leq UB_i; \; 0_i \leq e_j \leq \left[E_j\right] \text{ (eq. 4.6)}$$

Which allows over fluxes and enzyme usage reactions, the latter having an upper limit equal to the concentration of each measured enzyme in a proteomics dataset. For unmeasured enzymes, an additional constraint is imposed in the form of:

$$\sum_p Mw_j * e_j \leq \; P_{tot} * \sigma \text{ (eq. 4.7)}$$

Where P_{tot} is equal to the total protein mass available for expression of metabolic enzymes and σ accounts for an average saturation factor across all unmeasured enzymes.

4.2 Analysis of the kinetic parameters in the BRENDA database

The main incorporation of the GECKO modeling framework is enzyme constraints, based on turnover number and molecular weight of enzymes. The reconstruction of ecYeast7 required the formulation of a set of criteria for matching kinetic parameters to enzymes and reactions in the model. This classification was based on giving priority to parameter entries reported for *S. cerevisiae* and the specific substrates of the reactions in the model, but also allowing it to complement the parameterization relaxing the match at the organism, substrate, enzyme commission number, in a progressive way.

The first GECKO version matched more than 3,000 kinetic parameters to yeast7, 48% of them coming from entries reported for enzymes in other organisms. Moreover, for 56% of the introduced parameters it was necessary to relax the searches at the level of EC number, by progressive addition of wild-cards (e.g., EC1.1.1.X, EC1.1.X.X), due to lack of reported entries for all the reactions in the model. All these aspects raise concern for implementation of the method on GEMs for other less studied organisms, especially since quantitative prediction of ecGEMs is highly dependent on the selection of parameters[156,191].

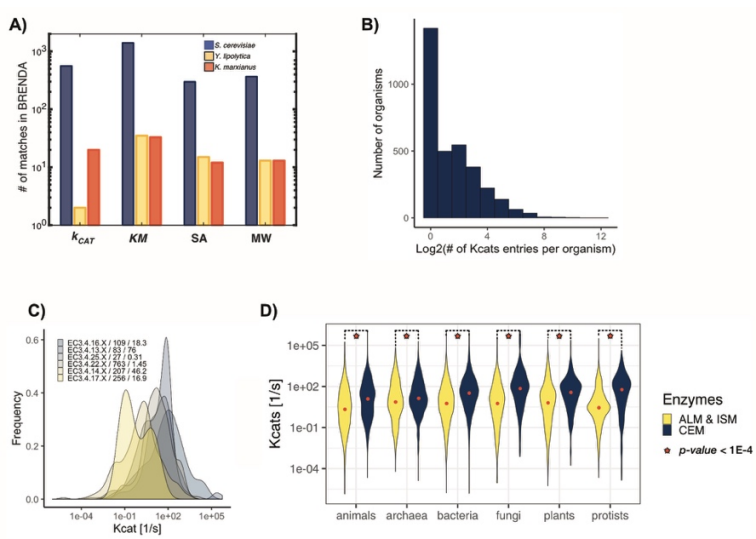


Figure 19.- A) Availability of kinetic parameters reported for *S. cerevisiae, Y. lipolytica* and *K. marxianus* in the BRENDA database (queried on 2017/05/25). k_{cat}). k_{cat} - enzyme turnover number, *Km*- Michaelis constant, SA- enzyme specific activity, MW- molecular weight. B) Distribution of the number of k_{cat} entries per organism in BRENDA. C) Distribution of reported k_{cat} values for different classes of EC3.4.X.X enzymes. D) Distribution of reported k_{cat} entries classified by phylogenetic origin and metabolic context (yellow color for amino acid and lipid metabolism, and intermediate and secondary metabolism; blue color for central carbon and energy metabolism). P-values compare yellow to blue distributions under a one-sided Kolmogorov-Smirnov test.

As the BRENDA database is the main source of kinetic parameters for GECKO, a global statistical analysis of the database was performed. All the available parameters, reported as k_{cat} or specific activities (k_{cat} =*SA*Mweight*), were retrieved from BRENDA, restricting the queries to non-mutant enzymes, in order to reflect their natural activity. This returned 39,280 parameter entries for 4,130 unique EC numbers, annotated with substrate name and organism of origin. Phylogenetic information for every organism was retrieved from the KEGG phylogenetic tree[192].

A first inspection showed a large bias of the compiled database towards few organisms, such as *S. cerevisiae*. **Figure 19A** shows that kinetic parameters are found in the order of hundreds or thousands, depending on the category, while very few parameters are reported for *Y. lipolytica* enzymes (just 2 k_{cat} values) ad *K. marxianus* (21 k_{cat} values). This range of number of entries is the same for most of the organisms with entries in the database, while just 5 model organisms account for the 24% of the total entries, as shown in **figure 20B**. It was also found that reported kinetic parameters are highly variable and may span several orders of magnitude even for closely related enzymes. Significant differences were found for distribution of k_{cat} values for closely related enzymes (classes with one or two imputed wild-cards), as the example provided in **Figure 19C**.

Classification of k_{cat} values by phylogeny and metabolic context (associating E.C. numbers to enzymes in the KEGG metabolic superpathways), revealed that central carbon and energy metabolism (CEM) enzymes tend to be significantly more active than those or in amino acid, lipid, intermediate and secondary metabolism, across the 5 KEGG kingdoms of life (animals, archaea, bacteria, fungi, plants and protists), **figure 19D**. Additional comparison of CEM enzymes divided by kingdoms of life showed that entries reported for fungi display significantly higher values, on average, than those for other kingdoms. The complete details of the statistical analysis and data processing of the BRENDA database are available in the supplementary material of **paper III.**

4.3 Development of GECKO 2.0: a toolbox for integration of kinetic and omics constraints into metabolic models

In **paper III**, the GECKO toolbox was updated and expanded to its 2.0 version following an open-source community development approach. This consists of making not just the code, but the whole history of development of the software, in which every new addition or removal to the code is justified by the author of the modification. This was adopted using a web-based git version control platform, enabling organized collaboration and programing by several users without the risk of damaging stable versions of the software.

Based on the findings of the previous section, the set of criteria for matching kinetic parameters to reactions in a GEM were modified for GECKO 2.0. In the new

algorithm, the lack of parameters for the modeled organism is circumvented by matching a parameter annotated to the phylogenetically closest organism with available k_{cat} for the same reaction. This assumption was inspired by the reduced rate of mutation in conserved genes across budding yeasts, such as metabolic genes, reported in **paper** I.

The introduction of wild-cards by the algorithm cannot be fully avoided when parameterizing a genome-scale model, as not all biochemical reactions have been characterized to the kinetic level, however, this problem is partially alleviated by the introduction of additional 8,118 new entries, which were found as specific activities in the database and were not part of the parameters pool in the first version of GECKO.

The code was refactored and adapted to a general structure of a GEM, extending its applicability to basically any GEM compatible with the COBRA or RAVEN toolboxes. In accordance with the claims in **paper II**, the output models from GECKO 2.0 are returned and stored in SBML L3V1 FBC2[186] for compatibility with any systems biology modeling or simulation software. Model functionality is verified, and overconstraining parameters are automatically flexibilized, in order to deliver an ecGEM that reproduces experimentally observed growth rates (provided by the user). Additionally, version control of the SBML file of a yeast ecGEM enabled tracking the effects that modifications in the software have on the ecGEM structure and basic functionality, contributing to reproducibility of results.

The model ecYeast7 was reconstructed using both versions of GECKO (1.0 and 2.0) to enable evaluation of the impact of the software upgrade over the model structure and performance. GECKO 2.0 proved to be effective at reducing the number of parameter matches with flexibilized EC number queries (introduction of wild-cards), from 1,817 to 556 (56% to 17% of the total number of parameters). The effect of the phylogenetic distance matching criterion was assessed by comparing the distribution of k_{cat} values retrieved by the two versions of ecYeast7 to the distribution of values reported for *S. cerevisiae* and all Fungi.

Statistical analysis showed that the algorithm in GECKO 2.0 is able to resemble the distribution of kinetic parameters for *S. cerevisiae* available in the data in a better way. Further comparison of batch growth rates on 19 different environments found a small increase in the average accuracy of predictions. Altogether, this demonstrates that GECKO 2.0 is effective at improving the specificity parameterization of a model, trying to resemble a phylogenetically related kinetic distribution without sacrificing predictive power.

Finally, computational times were drastically reduced, documentation extended and improved and a series of simulation utilities were added to the toolbox. The aim of the GECKO toolbox is to provide a software resource that can aid the research even outside the projects of our research group, it is a public resource.

4.4 Reconstruction of catalogue of ecGEMs for diverse organisms

The GECKO 2.0 toolbox was used to reconstruct ecGEMs for the three budding yeasts *S. cerevisiae*, *K. marxianus* and *Y. lipolytica*. To further test the functionality of the toolbox, the model *i*ML1515 for *E. coli*[193] and Human1 for *H. sapiens* metabolism[160] were extended with enzyme constraints. The size of the initial models and resulting ecGEMs is shown in **table 5**. The large increase in model size (number of reactions and metabolites) is introduced by the pseudo reactions and metabolites that balance the use of enzymes in the ecGEM[156]. Notably, GECKO 2.0 produced ecGEMs with an enzyme-gene coverage (percentage of metabolic genes with kinetic parameters) between 71-88%.

Nevertheless, metabolic modeling is a constantly evolving field. Additions and corrections are typical in the lifetime of a model or modeling software, therefore, changes in an ecGEM structure can be induced either by changes in the original model source, changes in the GECKO toolbox software, or changes in the software dependencies, such as RAVEN and SBML. Accounting for all these factors make version-control of a unified catalogue of ecGEMs into a cumbersome task. To solve this problem, an automated virtual continuous integration platform was developed and made publicly available, ecModels container.

Table 5.- Size metrics summary for the ecGEMs catalogue.

Original GEMs					
Organism	*S. cerevisiae*	*Y. lipolytica*	*K. marxianus*	*E. coli*	*H. sapiens*
Model ID	yeastGEM_8.3.3	*i*Yali4	*i*SM996	*i*ML1515	Human1
Reactions	3963	1924	1913	2711	13101
Metabolites	2691	1671	1531	1877	8400
Genes	1139	847	996	1516	3628
Enzyme constrained GEMs					
Model ID	ecYeastGEM	ec*i*Yali	ec*i*SM996	ec*i*ML1515	ecHumanGEM
Reactions	8028	3881	5334	6084	46259
Metabolites	4153	1880	2064	2334	12191
Enzymes	965	647	716	1259	3224
Enzyme coverage	84.72%	76.39%	71.89%	83.05%	88.86%
Reactions w/ k_{cat}	3771	1586	2891	2562	27014
Reactions w/ Isoenzymes	504	205	532	456	3791
Promiscuous Enzymes	572	324	469	673	2184
Enzyme complexes	252	75	27	383	756

This platform consists of an automated pipeline that constantly checks for version changes in the git repositories of the original model source, the GECKO toolbox,

and the RAVEN toolbox; if any change is detected the whole model reconstruction pipeline is run for all the models in the catalogue. In this way, all ecGEMs in the catalogue reflect the current version of the GECKO toolbox, and the current amount of knowledge accumulated in their sources, without the need of extensive curation or manual work.

The ecModels pipeline stores every model in each iteration using git and unique version identifiers, facilitating traceability of changes. All ecGEMs available in the catalogue are tested and calibrated by GECKO, except for ecHumanGEM which provides a knowledgebase of metabolism and kinetic parameters that is not specific to any kind of human cell or tissue, therefore, not suitable for simulation.

4.5 Testing functionality of automatically reconstructed ecGEMs

As pointed out before, one of the main advantages of extension of a GEM with enzyme constraints is the reduction of the solution space, discarding phenotypes that are not within the kinetic capabilities of the cell. Incorporation of kinetic constraints with GECKO 2.0 proved to be effective at reducing the solution space for the new ecGEMs for *K. marxianus*, *Y. lipolytica* and *E. coli*. Flux variability analysis (FVA) was run using the built-in FVA utility in GECKO 2.0 and compared to the variability ranges obtained for an equivalent GEM. Two different constraint scenarios were tested, substrate-limited (by fixing a low glucose uptake rate) and protein-limited (by enabling any substrate uptake rate and fixing growth rate to the μ_{max} of the ecGEM. This analysis revealed that enzyme constraints reduce the number of totally variable fluxes (reactions that can take any flux value between -1000 and 1000 mmol/gDw h, which is an undesirable model trait) to zero, in all organisms and conditions (**figure 20**).

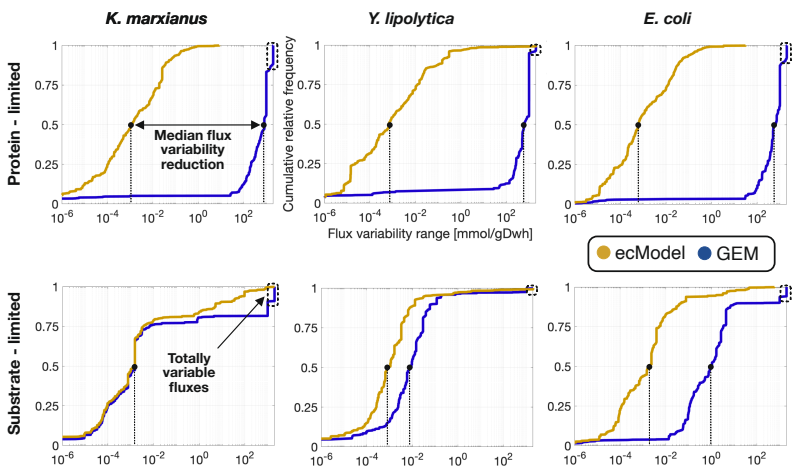


Figure 20.- Cumulative distributions for flux variability ranges predicted by GEMs and ecGEMs for *K. marxianus*, *Y. lipolytica* and *E. coli* under substrate-limited and protein-limited conditions.

Additionally, ecGEMs predicted flux variability ranges with significantly lower values than those in conventional GEMs, especially in protein-constrained regime, where the differences in median flux variability ranges can reach even several orders of magnitude. Notably, the median flux variability range displayed by ecGEMs did not show to be as sensitive to the magnitude of the carbon flux as conventional GEMs, and the predicted median variability range values are in the order of 10^{-4} to 10^{-3} mmol/gDw h, which may be negligible when compared to the precision of flux estimation by experimental approaches.

Prediction of cellular growth and total protein content of cells under diverse environments in *E. coli* by ec*i*ML1515 were compared to those of a metabolic and gene expression model (ME-model)[194]. Despite the high degree of detail regarding protein expression in ME-models, ec*i*ML1515 proved to improve prediction of cellular growth, while predicting cell protein contents within the range of predictions of the ME-model, demonstrating that ecGEMs are capable of capturing metabolism and cell physiology by using a simple description of kinetic limitations.

4.6 Evaluation of the impact of proteomics constraints on ecGEMs predictions

GECKO offered the first method for streamlined integration of proteomics measurements as constraints for metabolic networks. As enzymes are explicitly included as model components, and enzyme occupation by substrates is emulated by the introduction of enzyme usage reactions, which represent the concentration of an enzyme that is needed to sustain the flux through all the reactions that are catalyzed by it. Thus, absolute measurements (abundance data), in units of mmol/gDw can be applied as usage constraints for each of the enzymes in the model.

The proteomics data obtained from steady-state cultures of *S. cerevisiae*, *K. marxianus* and *Y. lipolytica* under environmental stress, collected and quantified for **paper I**, were reprocessed and transformed into absolute abundance values, by comparison of MS/MS spectra with data obtained from a calibrated external standard[77]. Enzyme abundances from standard and stress conditions were then integrated into the ecYeastGEM, eciSM996 and eciYali models, generating 11 condition-specific constrained models (4 for *S. cerevisiae* and *K. marxianus* and 3 for *Y. lipolytica*), aiming to gain more understanding of the stress responses of budding yeasts. The procedure for incorporation of enzyme abundance constraints in GECKO was revisited, systematized, and incorporated into the GECKO 2.0 simulation utilities. A description of the proteomics integration procedure is provided in the supplementary material of **paper III**.

To assess the effect of individual protein constraints on flux predictions, flux distributions were obtained for all organisms and conditions. These simulations were

obtained using FBA and, for each case, three distributions were obtained, one with ecGEM + proteomics, another one with the ecGEM without proteomics data (constrained by the total protein pool), and a distribution using the original GEM without kinetic parameters. To do so, different objective functions were used. For all cases flux constraints on measured glucose uptake were introduced. For the ecGEM with proteomics, FBA was run by minimizing the total utilization of unobserved proteins (non-measured); for the ecGEM without proteomics and the conventional GEM, an objective function that maximizes the amount of non-growth associated energy expenditure (NGAM) was used, as a way to emulate the additional nutrient demands of stressed cells[156].

Cumulative distributions of fluxes showed significant differences when comparing predictions between models with and without constraints, indicating a different allocation of flux when proteomics data are introduced. When focusing on prediction of enzyme demands between ecGEMs with and without proteomics constraints, most enzymes are predicted in a range between 0.5 and 2 fold-change ($FC = {}^{e_{ecm}}/_{e_{ecp}}$). Notwithstanding, it was found that, across organisms and conditions, 12-21% of utilized enzymes are predicted to be completely activated (going from 0 usage in non-proteomics model to a finite value in the data-constrained one) or deactivated (going from finite usage in non-proteomics model to a 0 value in the data-constrained one), most of them being in the former direction.

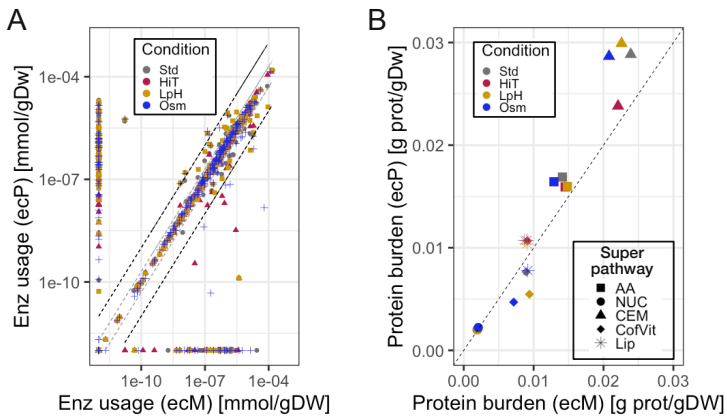


Figure 21.- Effect of protein constraints on ecYeastGEM predictions. **A)** Pairwise comparison of predicted enzyme usage between ecYeastGEM (ecM), and ecYeastGEM constrained with proteomics data (ecP). Std- standard condition, HiT- high temperature, LpH- low pH, Osm- high osmotic pressure. **B)** Pairwise comparison of total protein burden per superpathway predicted by ecYeastGEM (ecM), and ecYeastGEM constrained with proteomics data (ecP). AA- amino acids metabolism, NUC- nucleotide metabolism, CEM- central carbon and energy metabolism, CofVit- metabolism of cofactors and vitamins, Lip- lipid metabolism.

Predictions for *S. cerevisie* are shown in **figure 21A**, where the "activated" enzymes are those vertically aligned in the left-most part of the plot, and the "deactivated" ones are the points parallel to the x-axis in the lower part of the plot. These prediction patterns are mostly caused by a more diversified use of isoenzymes in the more constrained models, in contrast to the optimal solutions in which the most efficient isoform is active. This use of isoenzymes is also suggested by proteomics data in which usually several isoforms of the same enzyme can be found as expressed.

Protein resource allocation to different sectors of metabolism was computed as the total sum of predicted protein demands (in mass terms) for the enzymes present in metabolic superpathways, according to KEGG pathway classification[192]. The predicted total protein mass allocated to enzymes in central carbon and energy metabolism was affected by a 20-28% increase in 3 out of the 4 conditions (**figure 21B**), which is significant as this is the highest metabolic protein burden of the cell. Thus, the protein data and flux predictions suggest that, under these conditions, protein expression of this sector of metabolism does not follow an efficiency maximization pattern. Predictions for the rest of metabolic sectors showed considerable consistency between the two different levels of constraints, indicating a protein-efficient expression pattern in these sectors.

4.7 Identification of constraining enzymes in stress conditions

Analysis of proteomics data across condition can inform about presence and differential expression of individual proteins. By contextualization of these data into an ecGEM, flux predictions can provide an estimation of enzyme metabolic activity or (i.e. how much of an enzyme is used to carry a given reaction flux). Enzymes that showed a non-decreasing expression level, between stress and standard conditions, and an increased demand in predicted flux distributions were identified across all stress conditions and yeast species. Results were narrowed down by focusing just on those for which a relative enzyme usage (a proxy to its saturation) is equal or higher than 0.95 (highly saturated). These enzymes represent proteins that are not down-regulated by the cell, and are more demanded by the metabolic network under stress, up to the level of becoming potential flux limitations.

A total of 16 enzymes following this increased metabolic demand pattern were identified across all differential comparisons (stress vs standard), and are listed in **table 6**. Interestingly, none of the genes encoding for these enzymes were found to be significantly DE (in any direction) among the results reported in **paper I**. The fact that none of these enzymes seems to be limited by regulation at the transcript and protein level under stress conditions, parallel to their increased demand from the network, suggests them as candidate targets for engineering flux patterns. Consideration of the location in metabolism for some of these enzymes highlights their potential use for bioproduction, as several of them are either in or close to the TCA cycle, provider of building blocks, and amino acid metabolism.

Table 6.- Highly saturated enzymes with non-decreasing expression levels in budding yeasts under stress exposure. Single-copy orthologs across the three species are indicated in grey. Gene names starting with KLM correspond to *K. marxianus*, those with YALI to *Y. lipolytica*.

Enzymes	Genes	Short names	SubSystems	HiT	LpH	Osm
P37291	YLR058C	SHM2	Glycine, serine and threonine metabolism; Glyoxylate metabolism; Folate metabolism	X		
Q07500	YDL085W	NDE2		X		
P38858	YHR163W	SOL3	Pentose phosphate pathway		X	
P53315	YGR248W	SOL4	Pentose phosphate pathway		X	
P00958	YGR264C	MES1	Selenocompound metabolism; Aminoacyl-tRNA biosynthesis			X
P28777	YGL148W	ARO2	Aromatic amino acids biosynthesis			X
P32895	YKL181W	PRS1	Pentose phosphate pathway; Purine metabolism; Biosynthesis of amino acids	X		X
P52489	YOR347C	PYK2	Gluconeogenesis; Purine metabolism; Biosynthesis of amino acids	X		X
W0T7K6	KLMA_30312		Citrate cycle (TCA cycle); Propanoate metabolism		X	
W0TGN7	KLMA_70385		beta-Alanine metabolism		X	
W0TCW7	KLMA_60181		Arginine and proline metabolism	X	X	X
Q6BZU8	YALI0F30745g		Folate metabolism	X		NA
Q6C5P5	YALI0E16346g		Glycine, serine and threonine metabolism; Glyoxylate metabolism; Folate metabolism		X	NA
Q6C5R5	YALI0E15818g		Alkane metabolism		X	NA
Q6C6P0	YALI0E07766g		Alkane metabolism		X	NA
Q6CGX5	YALI0A15147g		Alkane metabolism		X	NA

Finally, genes encoding for the enzymes in the ecGEMs were mapped to the list of 2,959 shared single copy orthologs in the three species. Despite some of the enzymes shared by the three yeasts, being among the highly saturated enzymes in **table 6**, these seem to be responsive to stress in a single species. A less stringent search found that, from all the highly saturated enzymes, none of them showed to be highly saturated in more than one species. This finding is consistent with those in **paper I**, that say that each of these species has evolved different molecular mechanisms to tolerate environmental stress, which here is shown to be conserved at the level of enzyme limitations.

4.8 Summary

In **paper III** GECKO, a software toolbox for incorporation of kinetic and omics constraints into genome-scale metabolic models, was extended and upgraded to its 2.0 version. The software was generalized to be capable of processing any GEM structure with a format compatible with COBRA and RAVEN toolboxes. The number of kinetic parameters available for incorporation into model reactions was increased and the matching algorithm was improved, based in phylogenetic distances between organisms. These modifications increased the quality of model parameterization, resembling the kinetic profile of the modeled organism in a better way, without compromising prediction accuracy in ecYeastGEM. The toolbox is provided as an open-source software resource, open to community development.

Furthermore, a pipeline for reconstruction of ecGEMs was automated and connected to the original sources of GEMs *for S. cerevisiae, K. marxianus, Y. lipolytica, E. coli* and *H. sapiens*, thus, enabling the construction of a catalogue of continuously updated and version controlled ecGEMs for diverse budding yeasts and model organisms. Reduction of the solution space by incorporation of enzyme constraints proved to be efficient in the microbial models of this catalogue, even more for protein-limited conditions, which arise at high growth rates or high substrate availability.

Integration of proteomics data from stress condition samples into the ecGEMs for the yeast species in the catalogue predicted that, despite being phylogenetically related, these three species are faced with different enzymatic limitations in their metabolic networks when exposed to environmental stress. Nevertheless, comparison of these results with DE data enabled identification of a few potential gene targets for controlling flux under stress conditions that are also found in biotechnological applications, which may be of interest for metabolic engineering purposes.

Overall, this study provides the systems biology community with research infrastructure (software and models) and an example on how to integrate diverse omics data and metabolic networks for gaining biological insight that eludes other data analysis approaches.

5.- The use of enzyme constraints for rational systems metabolic engineering

5.1 Current constraint-based methods for prediction of metabolic engineering targets

Genome-scale metabolic models have found diverse uses in metabolic engineering and synthetic biology. GEMs constitute a knowledgebase on the metabolism of organisms, thus, they provide a highly detailed metabolic map for rational design of engineered strains. Several constraint-based methods are available for prediction of intervention strategies, that rewire metabolism towards production of compounds of practical interest.

The method optKnock, developed in 2003, which focuses on identification of fluxes that should be knocked-out to couple bioproduction and cellular growth. This approach has been successful for improving production of diverse chemicals in different hosts (e.g., biofuels precursors in *Bacilus subtilis*[195] and *S. cerevisiae* cells[196]). OptForce, a method that finds flux candidates for knock-out and also reaction fluxes that must be increased to couple production to growth, was developed in 2010 and rapidly gained popularity for model-driven metabolic engineering[197–202]. Other extensions to these methods, such as K-optForce[203], accounting for kinetic expressions for metabolic reactions, whenever known; and optGENE[204], which directly assesses gene engineering interventions instead of reaction fluxes, have also been developed.

A different approach, the flux-scanning with enforced objective function algorithm (FSEOF)[205], also developed in 2010, is based on the trade-off between optimal growth and increased chemical production by a metabolic network, thus, enabling identification of reactions with an increasing flux pattern when switching the cellular objective. As this method is based in a series of FBA problems, it provides a simple framework for understanding the metabolic context and effect of engineered fluxes towards the production goal. In contrast, the methods mentioned above rely on more complex mathematical formulations, such as MILP problems, or even combination of differential equation systems together with MILP, as done in K-optForce, which may hinder the process of gaining understanding of the metabolic network from the predicted reaction targets. Nevertheless, a common problem of the current methods for prediction of gene intervention targets is their need for either establishing the maximum number of prediction outputs, or for selecting an arbitrary number of gene modifications from large lists of predictions, which may not result in an optimally producing mutant strain.

The simple structure of ecGEMs, compatible with constraint-based methods provide an opportunity for development of novel constraint-based methods that account not just for flux redistribution, but also account for the impact of kinetic differences

among pathways, and are capable of improving prediction of gene engineering targets.

5.2 Model-aided engineering of *S. cerevisiae* cells for intracellular heme accumulation

In **paper IV** a metabolic model of yeast (yeast 7.6[183]), together with a modified implementation of the FSEOF method, were used for prediction of gene engineering targets to increase internal accumulation of heme, a cofactor essential for aerobic life and a crucial component of human hemoglobin, interesting for medical and food research purposes[206,207].

The FSEOF method consists of running a series of FBA problems, maximizing product yield, subject to decreasing levels of biomass yield (g biomass produced / g of carbon source). A flux score, representing a normalized slope of the flux change across simulations, is then assigned to each reaction. Traditionally, flux scores higher than unity identify candidate fluxes for amplification, however, the role of reaction fluxes with scores lower than one has not been extensively explored in FSEOF approaches. These decreasing fluxes indicate reactions that should carry less flux, in comparison to a wild-type, in order to increase the production of the desired metabolite, therefore, in this project predicted flux scores lower than one were considered as targets for reduction, and those with a zero score, as candidates for complete knock-out. Flux scores were transformed into gene scores by averaging the flux score across all reactions catalyzed by a given gene product. This approach is illustrated in **figure 22**.

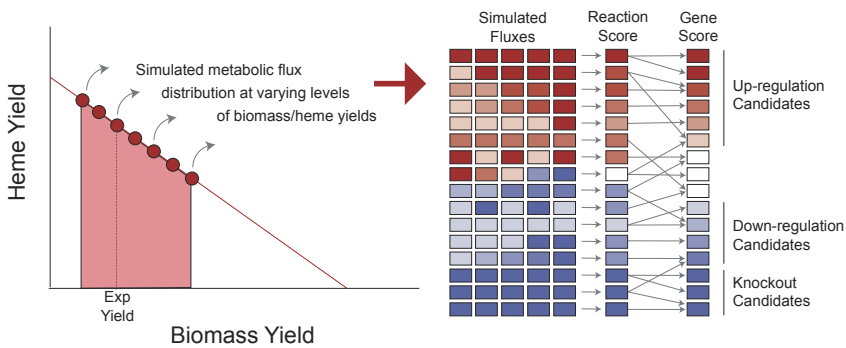


Figure 22.- Modified flux-scanning with enforced objective function algorithm. Flux scores are computed as the slope obtained between the initial and final flux for every reaction from a series of FBA simulations constrained by decreasing suboptimal biomass yields. Gene scores are obtained by averaging all flux scores of the reactions catalyzed by a given gene product. Gene scores higher than one indicate gene candidates for overexpression; scores between 0 and 1 indicate candidates for down-regulation (knock-down); and gene scores equal to 0 indicate gene candidates for deletion (knock-out).

The modified FSEOF approach predicted 84 gene targets for increasing accumulation of heme, from which 62 were classified as overexpression targets, 8 as deletions, and 14 as gene knock-downs. Gene candidates were experimentally modified in yeast cells (CEN. PK. 113-11c strain background). As gene down-regulation requires considerable fine-tunning in comparison to the other modifications, these gene targets were evaluated as gene deletions when implemented experimentally. From the 84 targets, 76 were successfully modified, individually, in the yeast cells. Intracellular heme concentration was measured from cell samples after 24 and 48 h of incubation. From the 15 tested deletions, 8 of them increased heme production; while from the 61 tested overexpressions, 32 proved to increase intracellular heme concentration.

Overall, validated gene targets encode for enzymes located in the heme biosynthetic pathway, glycolysis, pyruvate, Fe-S clusters, glycine, and succinyl-CoA metabolism. Just 4 out of the 40 successful modifications were capable of inducing an increase in heme levels higher than 150%. Notably, overexpression of HEM13 (encoding for coproporphyrinogen III oxidase, was found to be the most efficient modification, yielding a 300% increase in heme concentration. This catalogue of experimental results provides a systematic assessment of FSEOF predictions for a specific molecule in yeast, without imposing an arbitrary limit on the targets to test.

At the time in which the experimental work for construction and characterization of the 76 yeast mutants was finished, a new version of the yeast model was available[208], including extensive curation of metabolism, moreover, an ecGEM was constructed using the GECKO toolbox. The FSEOF-based approach from **figure 22** was used for predicting a new list of gene targets using ecYeastGEM, aiming to find a combination of predicted individual targets that could increase heme accumulation even further. This returned a list of 95 gene candidate targets.

ecGEMs enable direct simulation of gene deletions (by blocking the usage of their corresponding enzymes) and gene overexpressions (by allowing the model to increase the demand of an enzyme), and evaluation of their impact on the flux towards heme accumulation. Gene modifications that were detrimental for heme production in silico were discarded from the list of ecGEM predictions. After this, a total of 80 gene targets remained, from which 40 were found to have also been predicted by the initial model yeast 7.6. Inspection of the ecYeastGEM-exclusive predictions revealed a considerable number of gene targets in amino acid and pyruvate metabolism and the TCA cycle. Suggesting that the FSEOF-based approach together with an ecGEM can capture gene targets that contribute to rewire the wild-type biosynthetic flux patterns of the cell.

FVA applied to the enzyme usage reaction for the remaining 80 candidate targets enabled identification of redundant enzyme targets to reduce the number of candidates to 71. Furthermore, a sequential cumulative integration of the remaining gene targets into an in silico strain, allowed identification of a combination of 58

gene targets for optimal heme production that are predicted to be compatible, or suitable for combined implementation in yeast cells.

An optimized mutant strain was constructed following an iterative combinatorial approach, adding gene modifications one by one, informed by the predicted impact from in silico simulations and additional physiological/experimental criteria. Production performance, in terms of heme accumulation and cell growth, was evaluated after every inclusion of a new genetic modification. In cases in which an introduced modification resulted in a negative effect over heme production, then this was substituted by another one, however, discarded modifications were kept in the list of candidates for following iterations of the combinatorial approach.

This process acknowledges that the order of introduction of genetic modifications in a combined mutant influences the performance of the strain. As an example, deletion of GCV1 and GCV2 genes (encoding for subunits of the glycine cleavage complex) did not contribute to strain performance where implemented in combination with some other modifications, however they contributed to improve strain performance in further iterations of the combined mutant. This suggests that the metabolic flux patterns of a cell are a systems property, resulting from coordinated expression of all genes, and cannot be completely defined by modification of a single genes.

This sequential process, explained in detail in the supplementary information file of **paper IV**, produced a strain capable of accumulating 56 mg/L of heme intracellularly, representing a 70-fold increase in comparison to the initial strain, when normalized over the total produced biomass. The cumulative effect of introduced gene modifications is displayed in **figure 23A** in terms of measured intracellular heme concentration and observed optical density of the cell cultures (proportional to biomass concentration).

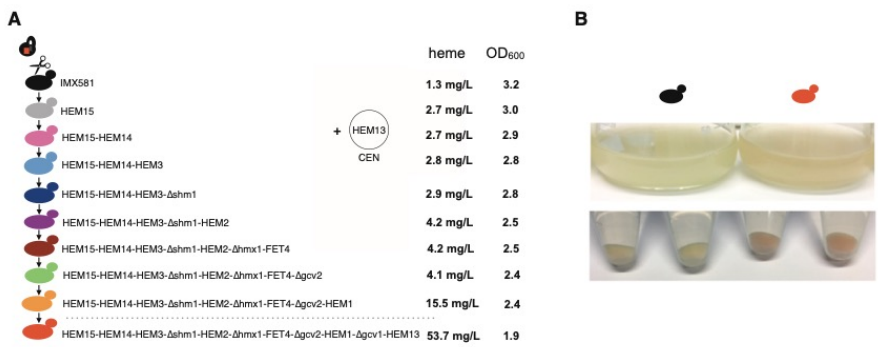


Figure 23.- Construction of a combined mutant *S. cerevisiae* strain improves intracellular heme concentration by 70-fold. A) Sequence of introduction of genetic modifications in the heme producing strain. Heme levels and cell culture OD_{600} were measured after 24 h of incubation. HEM13 gene was first introduced into the IMX581 S. cerevisiae strain using a CEN. PK. 113-11c centromeric plasmid. Gene modifications were enabled by integration of

CRISPR-Cas9 gene into IMX581 genome. Gene overexpressions were achieved by using expression cassettes carrying the *S. cerevisiae* TEF1 promoter, followed by the inserted gene and the ADH1 terminator. B) Cell cultures of the initial strain (black) and the final engineered strain (red) showed appreciable color differences after 24 h of cultivation. Color differences of cell extracts showed this pattern more clearly, with the mutant cells turning red, a sign of the increased intracellular heme concentration.

5.3 Some bloody lessons

The project described in the previous section offers an example of some intrinsic difficulties in metabolic engineering endeavors. Construction of mutant strains and phenotypic characterization are time-consuming tasks, that can prolong for months or even years. In the case of this project, while the initial sets of predictions were tested the yeast GEM underwent through major changes, and even a new framework in constraint-based modeling was developed, which change the course of the project.

From the initial set of experimental results, it was clear that not all predicted gene targets can induce the expected effects. This issue may relate to different aspects, in which model quality plays a decisive role, as errors in model components can result in false positive predictions. Additionally, GEMs are metabolic models that rely on stoichiometry constraints, ecGEMs add enzyme capacity to this, however the effects of complex kinetic mechanisms, dependent on metabolite concentrations, and those caused by the influence of regulatory gene networks over the state of proteins and enzymes, are not captured by these models, thus, apparent false positive predictions could find an explanation with further knowledge on cell physiology. This is case is shown for the case of individual overexpression of HEM4, which was found to be detrimental for the phenotype, possibly due to previously reported toxicity effects of uroporphyrinogen III, the product of the enzyme encoded by HEM4.

Interestingly, it was observed that several genetic modifications did not improve the cells performance significantly but proved to contribute to increased heme production when combined with others. Even more striking is the fact that the order of inclusion of genetic modifications in a multiple mutant strain has an impact on the cumulative effect of modifications. The example of GCV1 and GCV2 show how the potential of a genetic manipulation sometimes can be unleashed by modifications in other sectors of metabolism.

Explicit integration of enzymes into ecGEMs allows to consider the effects of gene redundancy in metabolism, thus, offering a platform for a more comprehensive search for an expression profile more suitable for chemical production. Additionally, their simple structure and the treatment of enzymes as pseudometabolites and pseudoreactions enables a straightforward simulation of genetic modifications. Nonetheless, as the approach taken in this project for evaluation of modifications in the ecGEM relied on comparison of optimal flux distributions (from FBA), quantitative computing of the effect of gene deletions under carbon-limited conditions remains challenging.

5.3 ecFactory: a method for refining prediction of genetic engineering targets for increased chemical production using enzyme constraints

The findings of **paper IV** suggest that ecGEMs are a suitable platform for refining predictions of gene engineering targets for increased bioproduction by constraint-based methods. The learnings from the iteration between computational and experimental work, explained in the previous section, motivated the development of a structured method that predicts an optimal combination of gene targets for increasing production, ecFactory, which is presented an applied at large-scale in **paper V**.

This method is rooted in an FSEOF-based approach for prediction of gene modifications for redirecting metabolic flux towards production of a desired metabolite. In summary, ecFactory consists of three major steps: 1) prediction of gene expression scores, indicating intensity and directionality of genetic modifications; 2) discard gene targets encoding for unfavorable enzymes (redundant, low efficiency) and; 3) Obtention of a minimal combination of modifications required for driving cells from optimal biomass formation to a metabolic production regime. The overall objective of this method is to reduce the number of predicted targets to an optimal metabolic engineering strategy, by taking enzyme allocation and connectivity into account.

FVA restricted to the enzyme usage pseudoreactions for the genes that are predicted by the initial step of the method, provides an estimate of the ranges of enzyme expression needed for achieving optimal production yield. In **figure 24A** a simplified representation of a metabolic pathway with enzymes is shown, in which 3 linear reaction steps are essential for the task of maximizing production of the final compound "D", additionally, and flux through a reaction that consumes the precursor of D for another purpose is detrimental for production. For the second reaction step there exist three different isoenzymes that can catalyze the reaction. **Figure 24B** displays the variability ranges for the enzyme demands in a pathway like this, comparison of these ranges with a parsimonious FBA solution, representing the minimal enzyme burden necessary for achieving a cellular objective, allows classification of enzyme in 4 classes: essential for production, optimal for production, suboptimal for production and futile for production. Enzymes that are either futile or suboptimal for production are discarded from the list of candidate targets. In the ecFactory pipeline, a suboptimal biomass production rate is fixed as a constraint, therefore, the enzyme usage variability ranges represent the demand of enzymes that can maximize production of the desired metabolite, while ensuring a fixed level of cellular growth.

As it was observed that for a given cellular objective, enzyme demand reactions can take variable flux levels, this variability should be considered for simulation of gene engineering for modification of enzyme expression levels. In the ecFactory method, a second FVA is run under the constraint of fixing growth rate to the maximum value possible for a unit carbon source uptake rate. Enzyme usage variability ranges

represent the expression levels expected from a "wild-type" strain that has evolved to maximize its growth yield. Comparison of the variability ranges for optimal production to those for optimal production (shown in **figure 25**) facilitate identification of gene target candidates encoding for enzymes with clearly differentiated expression patterns under the two different scenarios, which are prioritized in further steps of the algorithm.

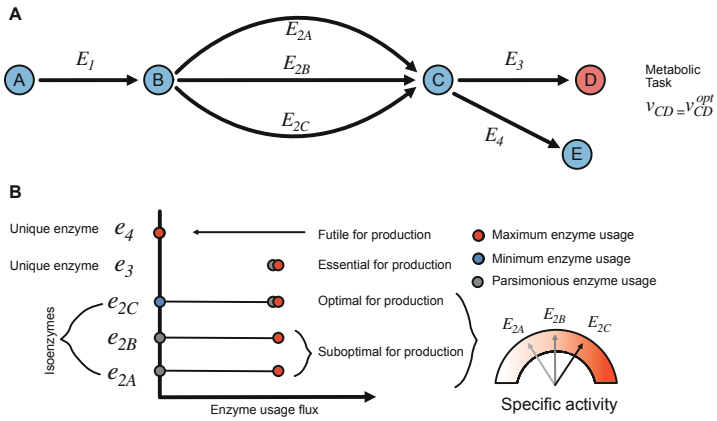


Figure 24.- Redundant enzymes in metabolism. A) A toy model network with the metabolic task of producing metabolite D from A. Unique enzymes catalyze reaction steps 1, 3 and 4, while reaction step 2 can be catalyzed by three different isoenzymes. B) Variability analysis on enzyme usage reactions. Blue and red points indicate the variability range for the usage of all enzymes. Grey points indicate the usage value obtained for each enzyme from an optimal flux distribution maximizing production of D. In FBA simulations, only the most efficient isoenzyme for a given reaction is the one carrying all the flux. Lowercase e_i indicate usage pseudoreaction for enzyme E_i.

After evaluation of enzyme usage variability ranges, the remaining gene target candidates for modification are incorporated into the ecGEM, one by one. Evaluation of production levels are performed by comparing maximum product formation rate and yield of individual mutants to those attainable by a wild-type strain, simulated by fixing the usage bounds of the enzyme candidates to those obtained from the variability analysis for an optimal growth scenario. Gene knockouts are simulated by blocking their corresponding enzyme usage reactions. For the case of gene knockdowns and overexpressions, the bounds of their respective enzyme usage reactions are changed from those obtained for the optimal growth scenario, to the usage bounds computed for an optimal production scenario in the variability analysis. Gene modifications that are detrimental for production are discarded from the list of remaining target candidates. The use of enzyme usage variability ranges for simulation of modification on enzyme expression levels is shown in **figure 26**.

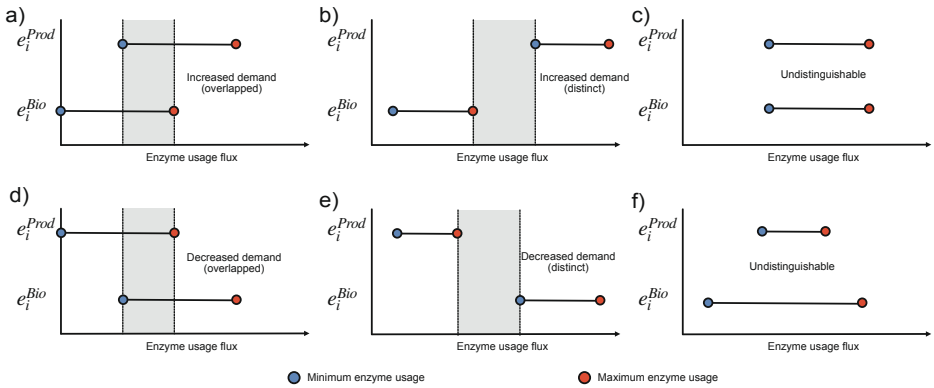


Figure 25.- Identification of optimal enzyme usage ranges under two different scenarios. e_i^{Prod} indicate enzyme usage in flux distributions maximizing for chemical production, e_i^{Bio} indicate enzyme usage in flux distributions maximizing for biomass production. A) Usage variability ranges for enzymes with increased demand under production and overlapping demand range with an optimal biomass scenario. B) Usage variability ranges for enzymes with increased demand under production and differentiated demand range between the two optimal scenarios. C) Usage variability ranges for enzymes with the same demand bounds predicted for both scenarios. D) Usage variability ranges for enzymes with the decreased demand under production and overlapping range with an optimal biomass scenario. E) Usage variability ranges for enzymes with decreased demand under production and differentiated demand range between the two optimal scenarios. F) Enzymes with undistinguishable demand ranges between the two scenarios, in which the production range is a subset of the enzyme usage range for optimal biomass production.

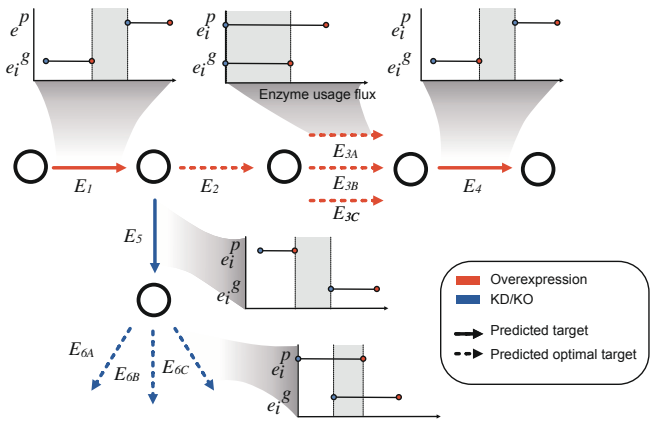


Figure 26.- Simulation of genetic modifications in a simplified metabolic pathway. Modification of enzyme expression levels is simulated by changing the bounds of each enzyme usage from those in an optimal growth scenario (e_i^g) to the ones computed for an optimal production scenario (e_i^p). KD.- gene knock-down, KO.- gene knock-out.

As a final step in ecFactory, a minimal combination of genetic modifications necessary to reach the predicted maximum production rate and yield is obtained. This is done by introducing all the remaining candidate gene modifications simultaneously in an ecGEM (i.e., fixing the usage bounds according to the ones predicted for an optimal production scenario for all enzymes encoded by the gene target candidates), maximum production rate and yield are then computed using FBA and a given suboptimal growth rate as constraint. Gene modifications are reverted, by changing the usage bounds of their enzymes back to the bounds predicted for a wild-type strain. If the removal of a genetic modification from the combined mutant strain does not affect in silico productivity levels, then the corresponding gene target is discarded from the list of candidates. Following this approach ensures that just the minimum set of modifications necessary for optimal production remain in the final list of targets.

5.4 Prediction of gene engineering targets for increasing production of 102 diverse chemicals in *S. cerevisiae* using the ecFactory method

The ecFactory method was used to predict gene targets for enhanced production of 102 different chemicals of industrial relevance, corresponding to diverse chemical families, in *S. cerevisiae* cells. This list of products is composed by 50 metabolites native to the *S. cerevisiae*'s network, while the other 52 are heterologous metabolites, present in other organisms. The classification of products by family is displayed in **figure 26B**. Heterologous production pathways were retrieved from the literature and incorporated into an ecGEM of *S. cerevisiae* (ecYeastGEM v8.3.4[157]). The method was successful at returning gene target predictions for all the 102 cases. Furthermore, the method proved to be effective at reducing the number of predictions in each of its sequential steps for all cases.

Global analysis on the number of targets per product revealed that step 2 in the ecFactory method (classification targets according to enzymatic characteristics, discarding redundant and suboptimal targets) is the main contributor to the total reduction in the number of predicted targets. Overall, the sequence of steps in ecFactory reduced the average number of predicted targets by 73%, from 85 initial targets (28 OEs, 42 KDs and 15 KOs) to 21 targets per product (7 OEs, 9 KDs and 5 KOs), as only optimal gene candidates, suitable for combination in a single strain, are kept in the final list of predictions.

After extensive literature review it was found that production of 22 of these products has been reported in engineered *S. cerevisiae* cells. Comparison of predicted targets with those implemented in each of these 22 cases found experimental validation for 28 different gene modifications across products. An interesting case was found in the predictions for increasing production of the alcohol 2-phenylethanol (used as a floral fragrance substitute), for which 7 out of the 12 predicted gene targets have been successfully implemented in highly producing strains of *S. cerevisiae*[209], *Y. lipolytica*[210] and *K. marxianus*[211].

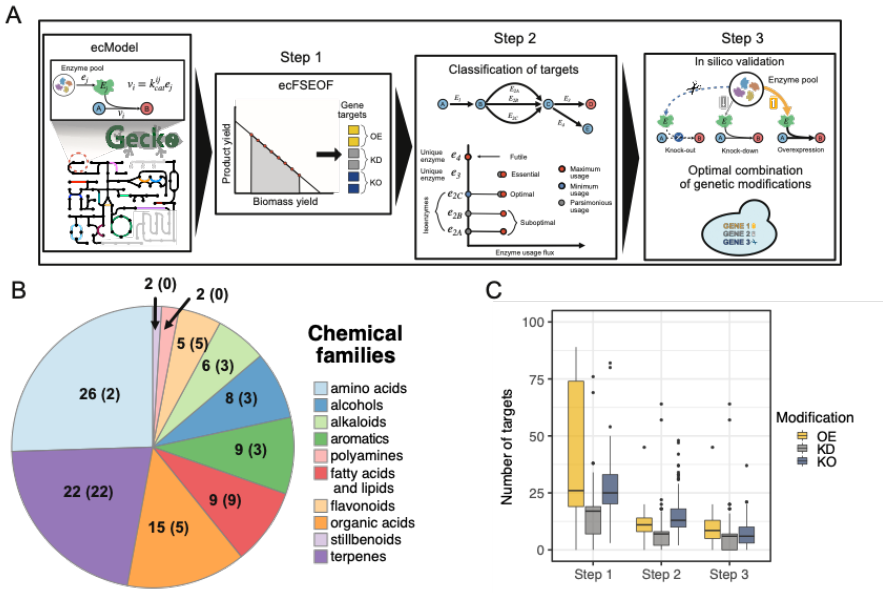


Figure 27.- Prediction of gene engineering targets for increasing production of 102 diverse chemicals in yeast using ecFactory. A) The ecFactory method and its three steps for prediction of gene engineering targets. B) Classification of 102 chemical products into 10 different chemical families. Numbers in each slice indicate the total number of products within a given family, numbers in parentheses indicate the amount of heterologous products in each family. C) Distribution of the number of predicted gene targets across the 102 chemicals after every step of the ecFactory method. OE.- overexpression, KD.- knock-down, KO.- knock-out.

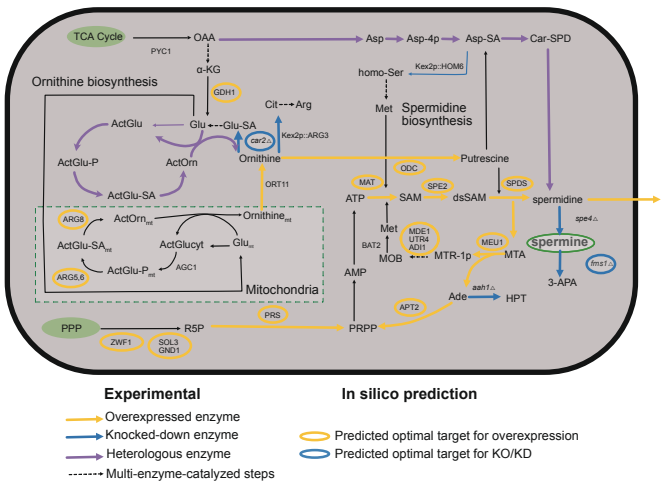


Figure 28.- Comparison between gene modifications implemented in vivo spermidine and ecFactory predictions for increasing spermine in *S. cerevisiae*.

Predictions of the ecFactory method were also able to capture 9 of the genetic modifications (overexpression of MAT, ODC, SPE2, SPDS, MEU1 APT2 and PRS, and deletion of CAR2 and FMS1) implemented in an engineered strain capable of producing high concentrations of spermidine in fed-batch cultures[212]. Engineering of this strain involved the coordination of modifications in several metabolic pathways, involving deletions, knock-downs and overexpression of native genes, and expression of heterologous pathways. **Figure 28** shows a comparison between experimental gene modifications and predicted targets by ecFactory for this case. Interestingly, predicted modifications resemble the general strategy followed by the experimental study, consisting of overexpression of the ornithine cycle, a direct precursor, together with the Yang cycle and some steps in the pentose phosphate pathway (PPP) to increase S-adenosyl-L-methionine, another important precursor of polyamines.

5.5 Evaluating the effects of enzyme capacity on bioproduction

The production envelope for each of the 102 chemicals was computed using both YeastGEM and ecYeastGEM under two different levels of constraining, low and high glucose uptake rate (1 and 10 mmol/gDw h, respectively). Overall, it was observed that the predicted envelop by YeastGEM is independent of the glucose uptake for all cases, as expected from a purely stoichiometric model (shown by purple and yellow dotted lines in **figure 29A**).

On the other side, production envelopes predicted by ecYeastGEM showed notable differences between high and low glucose uptake regimes for many of the modeled products (predictions for choline are shown as an example of this in **figure 29B**). In general, production yields for metabolites and biomass are negatively affected by high glucose uptake rates due to the tradeoff between substrate and protein utilization efficiencies characteristic of *S. cerevisiae*, which is caused by the kinetic differences between respiration (high protein burden and substrate efficiency) and fermentation (low protein burden and substrate efficiency).

For a large subset of products, it was found that in low glucose uptake regime, the optimal line of the production envelope predicted by yeastGEM and ecYeastGEM coincided in a region limited to the highest values of biomass yield, but after a critical point the slope of the optimal line of ecYeastGEM decreases, creating an enzymatically unfeasible region (shown in grey in **figure 29A**. This effect was predicted mostly for heterologous products, which tend to be produced by pathways catalyzed by enzymes with low efficiency. The list of heterologous products is mostly composed by terpenes, aromatic compounds, alkaloids, and flavonoids, which are naturally produced by plants in very low concentrations and correspond to pathways outside of central carbon metabolism. Consequently, as biomass yield is compromised in order to increase production flux of heterologous metabolites, the total enzyme burden increases until a point in which the cell switches to mixed respiro-fermentative metabolism, protein efficient but less substrate efficient than pure respiration.

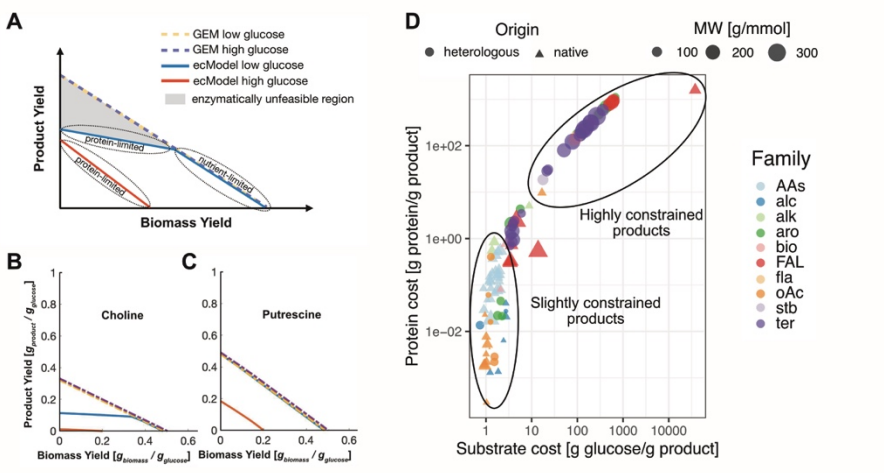


Figure 29.- Impact of enzyme capacity over production capabilities. A) Production envelope typical for a metabolite in a highly protein-constrained pathway. B) Production envelope for choline predicted by ecYeastGEM. C) Production envelope for putrescine predicted by ecYeastGEM. D) The landscape of production cost (substrate and protein cost) for chemical production in *S. cerevisiae* predicted by ecYeastGEM. AAs.- amino acids, alc.- alcohols, alk.- alkaloids, aro.- aromatic compounds, bio.- bioamines, FAL.- fatty acids and lipids, fla.- flavonoids, oAc.- organic acids, stb.- stilbenoids, ter.- terpenes.

FBA simulations were used to calculate the minimal substrate and protein cost of production for each the 102 chemical products (i.e., how much substrate or intracellular protein mass is needed for production of a gram of product). For this, a low glucose uptake rate (1 mmol/gDw h) and a null growth rate were used as constraints, assuming that the whole amount of substrate could be converted into the product of interest. Two well differentiated groups of products were identified among these predictions (shown in **figure 29D**), those with low substrate (<10 $g_{glucose}/g_{product}$) and protein (<1 $g_{protein}/g_{product}$) costs, and those with the opposite trend (>10 $g_{glucose}/g_{product}$ and (>10 $g_{protein}/g_{product}$).

FBA simulations showed that for all products of the high costs group, the flux of the glucose uptake reaction step did not reach its upper bound of 1 mmol/gDw h, however, the total demand of protein mass by metabolic enzyme was equal to the upper bound in ecYeastGEM. This means that the metabolic burden of the production pathways of these products is extremely high, so that the total mass of protein available for metabolic enzymes becomes the limiting resource, even at low glucose uptake rates. Furthermore, as increased protein demand of the production pathways demands additional substrate flux, due to the switch to mixed metabolism, the substrate cost for production of this group of compounds is increased beyond stoichiometric demands (i.e., production through pathways with very high protein demands induces additional substrate cost of production). The group of highly

constrained products is composed by 40 heterologous products and 5 native compounds in yeast, and especially enriched for terpenes. In contrast, the group of slightly constrained products (those with the lowest substrate and protein costs for production) was found to be composed mostly by native compounds (amino acids and organic acids).

5.6 Common gene modifications across multiple products suggest the possibility of platform strains for diverse chemical production

Genes predicted as candidate targets for more than one product were found among the 28 targets with experimental validation, mentioned in **section 5.4**, in particular, genes from the mevalonate pathway were predicted as targets for overexpression for increasing production of 9 different terpene compounds (table 1 in **paper V**). However, no common targets were found to be predicted for all of the 18 terpenes in this study, furthermore, the same case was found across all chemical families, except for flavonoids (catechin, genistein, naringenin, kaempferol and quercetin) for which 19 gene targets were identified as common to all these products (RNR4, RNR3, FAA1, ADO1, ARG5,6, CAR2, SAH1, ATP19, PPA2, RNR1, FDH1, IDP1, LPD1and MAE1, predicted as KO/KD targets, and MET6, FAA4, MDH2, ARG7 and ARG8 for overexpression). This result indicates that construction of a combined mutant, predicted to increase production for all 5 flavonoids, might be attainable. Nonetheless, this is an expected result, as flavonoids explored in this study come all from the same pathway, therefore the same metabolic rewiring strategy can provide the necessary precursor and cofactor demands for all these products.

In order to systematize the search for groups of products with shared gene targets, predictions were represented as a gene-product matrix, in which rows represent all the native genes in the *S. cerevisiae* network and columns represent the expression profile predicted for each product. Genes that are not predicted as targets for a product were assigned with a zero value, while KD targets as 0.25 (median gene score for KDs across all products) and OE targets a value of 4 (median gene score for OE targets across all products). Then, a series of t-SNE (t-distributed stochastic neighbor embedding) projections[213], exploring all the range of allowable hyperparameter values (perplexity). This series of data projections revealed 8 clusters of products with a tendency to group together despite the perplexity value, indicating high similarity in their predicted targets, this process of data representation is shown in **figure 30**. Furthermore, common targets were found for all clusters and are shown in **table 7**.

The clustering of products, according to their predicted targets, showed that for some cases, products cluster together because they are produced by the same pathways, or need the same metabolic precursors (clusters 5, 6, 7 and 8), nevertheless, clusters composed by compounds that are produced by very different pathways (clusters 1 and 2) where identified by this approach.

Table 7.- Common targets, per modification type, found across products in each of the identified clusters.

Cluster	Chemical Products	Shared KO targets	Shared KD targets	Shared OE targets
1	betaxanthin, caffeic acid, vanillin β-glucoside, β-ionone, glycyrrhetinic acid, miltiradiene, lycopene, taxadien-α-yl acetate, protopanaxadiol, genistein, quercetin, catechin, kaempferol, patchoulol, oleanolate, lupeol	RNR1, RNR4, RNR3, CAR2. FAA4, FAA1, FDH1	SAH1, ARG5,6, MET6, LPD1, ADO1, MAE1, ARG7, MDH2, ARG8, ATP19	NA
2	β-carotene, cinnamoyltropine, ARA, DHA, EPA, astaxanthin, psilocybin, docosanol	RNR1, RNR4, RNR3, CAR2. FAA4, FAA1, FDH1	IDP1, ARG5,6, LPD1, MAE1, MDH1, ARG7, PPA2, MDH2, ARG8, ATP19	NA
3	ergosterol, squalene, santalene, farnesene, amorphadiene, limonene, geraniol, artemisinic acid	NA	LPP1	PDB1, PDA1, PDX1, ERG12, ERG8, LAT1, MVD1
4	Itaconic acid, glutamine, proline, putrescine, spermine	NA	LPP1	PDB1, PDA1, PDX1, LAT1

5	valencene, nootkatone, linalool, β-amyrin	NA	ARG5,6, ARG8	ERG12, ERG8, MVD1
6	tryptophan, adipic acid, cis-muconate, hydroxymandelic acid	MAE1	LPP1	ARO4
7	phenylalanine, 2-phenylethanol, mandelic acid, cinnamate	MAE1	LPP1	ARO4, ARO1, ARO2, SOL3, GND1, ZWF1, PHA2, ARO7
8	Free-fatty acids, oleate, palmitoleate	NA	LPP1, ARG5,6, MAE1, CAR2, ARG8	CDC19, BPL1, SOL3, GND1, PDC1, ACS2, PPA2, ZWF1, ACC1, ALD6

Further analysis revealed that clusters 1 and 2 are mostly composed by heterologous products, identified as highly-protein constrained, due to the high enzymatic demands of their final production pathways. A consequence of this is that several genes involved in nucleotide metabolism and fatty acid metabolism are predicted as common targets for deletion, together with shared gene targets for KD in amino acids metabolism and TCA cycle, indicating that for increased production of these products, a compromise with the robustness of biomass formation pathways needs to be induced through rewiring.

Analysis of the demands of metabolic precursors and cofactors also revealed that increased demand of NADPH has an influence in the observed grouping of gene target profiles, as shown by the predicted OE targets in initial steps of pentose phosphate pathway (ZWF1, GND1 and SOL3) common to products in clusters 7 and 8.

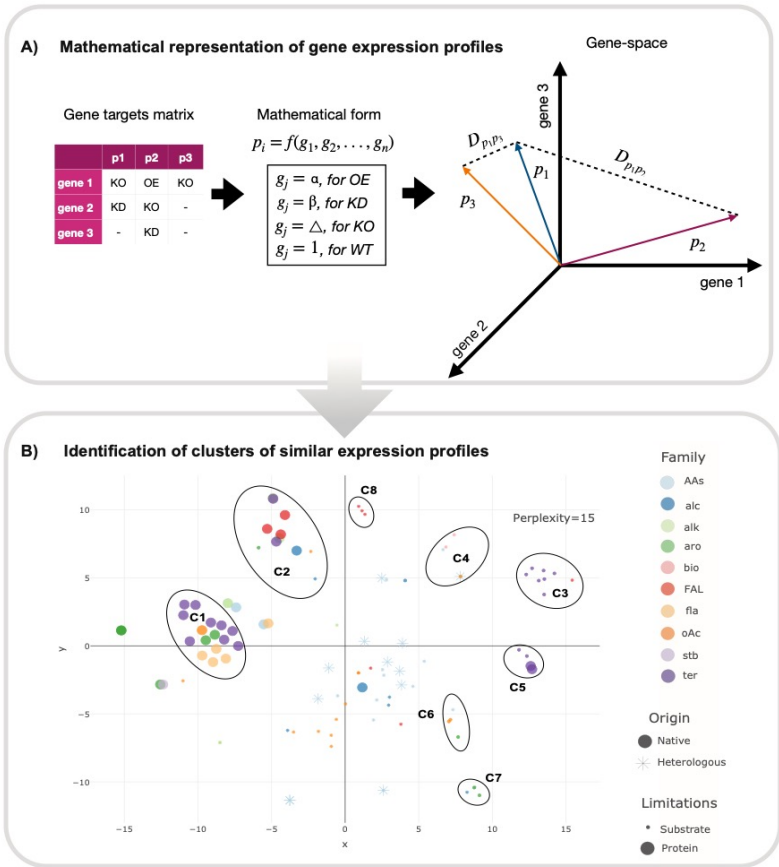


Figure 30.- Identification of similar predicted gene expression profiles across products. A) Gene target predictions were transformed into gene expression profiles for all products, the set of all predicted expression vectors formed a matrix suitable for numerical analysis. B) Iterations of t-SNE projections, spanning all allowable perplexity values and subject to 10,000 iterations per step, enabled visual identification of 8 groups of products with shared targets.

5.8 Summary

In paper IV, a GEM for yeast metabolism was used to predict gene engineering targets to increase intracellular accumulation of heme. 76 gene modifications were implemented independently in *S. cerevisiae* strain, from which 40 caused clear increased heme production. Then, an enzyme-constrained model of yeast was used to refine predictions and aid in the process of discarding gene targets according to the kinetic characteristics of their corresponding enzymes. An approach for identifying gene modifications that can be combined in a viable mutant was

developed and predictions guided the construction of a mutant strains, including 11 genetic modifications, that displayed the capability causing a 70-fold increase in heme accumulation in comparison to a reference strain.

Based on the findings in paper IV, a structured method for prediction of a minimal combination of gene targets for reaching optimal production levels of a metabolite of interest, using ecGEMs, was developed. This approach is based on a combination of the FSEOF method, FVA applied on enzyme usage reactions and FBA simulations to test the effect of modified enzyme expression ranges in the model. The method proved to be effective for drastic reduction of the targets initially predicted by FSEOF across 102 different products in *S. cerevisiae* cells.

A catalogue of metabolic engineering strategies for increased production of these 102 diverse chemicals was produced and made publicly available in **paper V**. Experimental validation for 28 of these targets across 22 different products was found in the literature, moreover, comparison of the predicted metabolic strategies for 2-phenylethanol and spermidine production were compared to previously reported experimental studies in which complex strategies were developed rationally. Predictions showed to capture the essence of the rationally engineered strains, and showed that our method can pinpoint targets that induce a complex rewiring of metabolism, involving up and down regulation of multiple pathways.

Finally, global analysis of the gene target profiles across the 102 products revealed the existence of multiple gene targets common to 8 specific groups of products. This finding suggests the utility of this approach for finding sets of genetic modifications that can be beneficial for the production of the necessary precursors and cofactors for groups of products, instead of designing product-specific strains. This approach can be leveraged for model-driven design of platform or chassis strains for diversified chemical production and accelerate the DBTL cycle in metabolic engineering and synthetic biology, both for industrial and scientific purposes.

Further analysis enabled identification of three basic factors that dictate the choice of products to be produced by a platform strain: the need for common metabolic precursors, the protein burden of the final production pathways, and the NADPH demands of the products aimed to be produced. These factors indicate that metabolic flux towards a given product of interest is an emerging systems property, in which many cellular processes are involved, and that optimal metabolic engineering strategies should be able to rewire metabolism in global way, finding the enzyme nodes that can cause the optimal way of pushing a cell from an optimal grower to an optimal producer.

6.- Understanding regulation of metabolism beyond enzyme capacity

To the memory of Stefan Hohmann.

Incorporation of proteomics abundance data into ecGEMs as constraints for the usage of individual enzymes enabled exploration of the impact of different environmental conditions over the metabolism of budding yeasts in **paper III**. Flux distributions across conditions and species revealed differences in predictions from ecGEMs with and without proteomics constraints, at the pathway, reaction, and enzyme usage level.

Flux distributions predicted without proteomics data represent the optimal flux and protein allocation profiles. However, it was shown that incorporation of data constraints induces expression and flux patterns that deviate from optimality. A limitation for the study of intracellular responses to changes in the environment with ecGEMs and proteomics data, is that the protein expression profile used as constraints is a systems property grounded in lower levels of causation, as transcriptional and translational processes regulate the state and expression level of proteins and metabolic enzymes. Therefore, a complete picture of the relation between the environment, the genotype, and the phenotype, remains elusive.

Multiscale models, such as ME-models (metabolic and expression models)[194,214] and resource balance analysis (RBA)[215,216] offer a representation of the processes involved in expression of proteins and their connection to metabolic reactions. Nonetheless, accuracy of phenotype predictions is usually compromised due to the hundreds, even thousands of parameters needed to construct them. Furthermore, some of these parameters represent quantities that cannot be directly measured.

Gene regulatory networks have been used for understanding the cascade of events that derive on control of transcription factors, and ultimately in the expression profile[217,218]. These models are usually represented as a network of Boolean interactions; thus, predictions are binary states of proteins and/or genes. Combination of Boolean models of gene regulatory networks and stoichiometric models have been implemented to refine the description of transcriptional regulatory mechanisms in *S. cerevisiae* cells, and for understanding the effects of hypoxia on the metabolism of Alzheimer's disease[219], among other applications. This hybrid approach enables study of the modulation of metabolic flux in cells under environmental or genetic perturbations.

In **paper VI**, a Boolean model of gene regulation induced by nutrient signaling is combined with an enzyme-constrained model of central carbon metabolism of *S. cerevisiae* cells. The hybrid model is used to explore and understand the cascade of

regulatory events that modulate expression of enzymes due to changes in the availability of nutrients.

6.1 Connection of an enzyme-constrained model with a gene regulatory network

For this study, a model of central carbon and energy metabolism, accounting for glycolysis, PPP, TCA cycle, oxidative phosphorylation, galactose metabolism and anapletoric pathways[154], was curated and extended with enzyme-constraints using the GECKO toolbox[157]. Using literature data available on gene regulation and nutrient-signaling in *S. cerevisiae*, a Boolean model for the main signaling pathways for carbon (PKA and SNF1) and nitrogen (mTOR), including crosstalk mechanisms between them, was also constructed.

Glucose uptake rate is used as an upper bound constraint in the metabolic model to predict a reference flux distribution, applying bi-level optimization, maximizing for growth rate, and then minimizing the total protein demand of the metabolic network. Assuming that glucose uptake by the cell is proportional to its extracellular concentration, the glucose uptake rate is used as a proxy to impose a Boolean condition of low (0) or high (1) availability of glucose. A threshold value indicating high glucose is then dictated by the uptake rate predicted at the critical dilution rate of *S. cerevisiae* (i.e., growth rate at which yeast cells switch for pure respiration to a respiro-fermentative metabolism), corresponding to a value of 3.29 mmol/gDw h. Glucose availability (0 or 1) is fed into the Boolean regulatory network, which runs a series of synchronous updates until steady-state is reached. As an output, the regulatory network indicates enzymes that should be down or up regulated. This transcriptional regulation response is represented in the enzyme-constrained model as follows. For upregulations:

$$lb_{E_i}^{reg} = e_i^{opt} + \rho\left(e_i^{max} - e_i^{min}\right) \quad \text{(eq. 6.1)}$$

And for downregulations:

$$ub_{E_i}^{reg} = e_i^{opt} - \rho\left(e_i^{max} - e_i^{min}\right) \quad \text{(eq. 6.2)}$$

Where $ub_{e_i}^{reg}$ and $lb_{e_i}^{reg}$ represent the regulated upper and lower bounds for enzyme *i*; e_i^{opt} is the enzyme usage value obtained for enzyme *i* in the reference optimal flux distribution; ρ is a regulation factor (>0) and assumed to be the same across all enzymes; $\left(e_i^{max} - e_i^{min}\right)$, indicates the variability range of the usage reaction for enzyme *i*, under the same conditions used to compute the reference flux distribution.

After imposing these constraints, a new FBA problem is solved to obtain a flux distribution, once again maximizing for cellular growth, and relaxing the constraint on glucose uptake rate by 15% (which corresponds to the maximum error of

predicted glucose uptake rate under carbon-limited conditions, and in the range of 0-0.4 h^{-1} of dilution rate, using a purely enzyme-constrained metabolic model of yeast[156]). Overall, this process assumes that the nutrient induced regulatory network modulates the expression of enzymes around a state of optimal allocation, enabling enzymes to be expressed outside of their optimal levels if required.

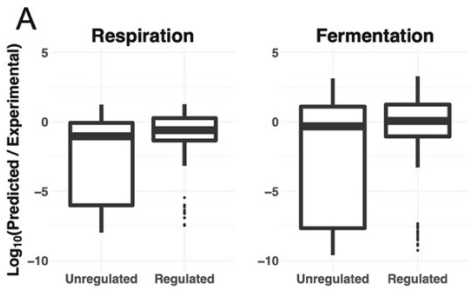
6.2 The impact of regulation over predictions of enzyme demands

The simulation setup described above was used to generate flux distributions of S. cerevisiae's metabolism in the whole range from 0 to its maximum specific growth rate, using glucose as a carbon source. Predicted exchange fluxes of glucose, oxygen, and the byproducts CO_2, ethanol, acetate, were compared against experimental measurements from chemostat cultures, over the whole range of dilution rates (0-4h^{-1}). Accuracy of exchange fluxes showed a median relative error of 9.82% across all exchanged compounds and dilution rates. Moreover, the hybrid model was able to predict the emergence of the Crabtree effect, switch from pure respiration to respiro-fermentative metabolism, at a critical dilution rate of 0.285h^{-1}, showing consistency with previous enzyme-constrained models and experimental measurements[154,156,220].

Due to its enzyme-constrained module, the hybrid model is also capable of predicting enzyme demands. These values were compared with protein abundance data from two independent studies in *S. cerevisiae*, one characterizing physiology under respiratory conditions (chemostat cultures at 0.1 h^{-1}) and another one at high growth rates and mixed metabolism (batch cultures at 0.4 h^{-1} growth rate). Log-10 transformed ratios between predicted demand of enzymes over the measured abundance (both in mmol/gDw h) were used as an error metric for comparison of each protein. This metric computes the error of prediction in orders of magnitude, in which the sign indicates under predicted (negative values) and overpredicted values (positive values).

Figure 31A shows a significant decrease in the range of errors for predicted protein demands by incorporation of the regulatory layer into the ecModel. This error metric was reduced from 2.62 to 1.55 in respiration, and from 3.56 to 2.32 in mixed metabolism. Furthermore, in respiratory conditions, 40.83% of the enzymes in the hybrid model were predicted in the same order of magnitude as their experimental counterpart. This percentage rose to 65.51% for the fermentative conditions. The distributions of log-10 transformed ratio values (**Figure 31A)** showed that a pure ecModel tends to underpredict enzyme usages for a significantly larger number of enzymes, in contrast to the hybrid model. Further analysis of this trend revealed that the metabolic model predicts a zero usage for many enzymes in the model, which display non-zero abundance values in the experimental datasets. This is a characteristic feature of protein abundance predictions with enzyme-constrained models and FBA, as an optimal flux distribution uses just the optimal isoform for carrying the entire flux of reactions with enzymatic redundancy.

Figure 31C displays a qualitative comparison between protein expression between the enzyme-constrained model, the hybrid model, and experimental measurements for each protein in the models. It can be seen how the cell tends to overexpress multiple isoenzymes for its reactions with redundancy, especially under respiro-fermentative metabolism. The hybrid model was successful at predicting a diversified usage of isoenzymes. In respiratory conditions, the regulation layer improved the prediction of diversified isoenzyme expression in several steps of glycolysis and oxidative phosphorylation. In mixed metabolism, this effect was observed in predictions for enzyme usage across the whole network. Notably, the hybrid model was successful at predicting usage of the 4 different glucose transporters in the model under mixed metabolism, which is consistent with the data.



B

	Respiration		Fermentation	
Metric	ecModel	hybrid	ecModel	hybrid
Sensitivity	0.38	0.66	0.33	0.65
Precision	0.82	0.60	1	1
Accuracy	0.55	0.5	0.35	0.65
FMI	0.57	0.63	0.58	0.80

C

rxn	ORF	genes	Resp.	Ferm.
ACS	YAL054C	ACS1	○ □ ▲	● □ △
	YLR153C	ACS2	● ■ ▲	● ■ ▲
ADH1	YBR145W	ADH5	● □ △	● □ △
	YOL086C	ADH1	○ □ ▲	● ■ ▲
ADK1	YDL166C	FAP7	● ■ ▲	● ■ ▲
	YDR226W	ADK1	● □ △	● □ △
ALD2	YMR110C	HFD1	● □ ▲	● □ ▲
	YMR170C	ALD2	○ □ ▲	● □ △
	YER073W	ALD5	● □ ▲	● □ ▲
	YOR374W	ALD4	● □ ▲	● □ △
CDC	YAL038W	CDC19	● ■ ▲	● ■ ▲
	YOR347C	PYK2	● □ ▲	● □ △
CIT	YNR001C	CIT1	○ ■ ▲	● ■ ▲
	YPR001W	CIT3	○ □ ▲	● □ △
COX1	COMPLEX I	COMPLEX I	● ■ ▲	● ■ ▲
	COMPLEX II	COMPLEX II	● □ ▲	● □ ▲
DAR	YDL022W	GPD1	● □ ▲	● □ ▲
	YOL059W	GPD2	● □ △	● □ △
ENO	YPL281C	ERR2	○ □ ▲	○ □ △
	YGR254W	ENO1	● □ ▲	● □ ▲
	YHR174W	ENO2	○ □ ▲	● □ ▲
	YMR323W	ERR3	○ ■ ▲	● ■ ▲
	YOR393W	ERR1	○ □ ▲	● □ ▲
GLD	YGR192C	TDH3	● ■ ▲	● ■ ▲
	YJL052W	TDH1	● □ ▲	● □ ▲
	YJR009C	TDH2	● □ △	● □ ▲
GND	YGR256W	GND2	○ □ ▲	● □ ▲
	YHR183W	GND1	○ ■ ▲	● ■ ▲
GPP	YER062C	GPP2	● □ ▲	● □ △
	YIL053W	GPP1	● □ △	● □ ▲

rxn	ORF	genes	Resp.	Ferm.
HXK	YLR446W	YLR446W	○ □ △	● □ ▲
	YCL040W	GLK1	● □ ▲	● □ ▲
	YFR053C	HXK1	● ■ ▲	● ■ ▲
	YGL253W	HXK2	● □ △	● □ ▲
MLS1	YIR031C	DAL7	○ □ △	● □ △
	YNL117W	MLS1	● ■ ▲	● ■ ▲
NDE2	YDL085W	NDE2	○ □ ▲	● □ △
	YMR145C	NDE1	● □ △	● □ △
PDC	YGR087C	PDC6	○ □ ▲	● □ ▲
	YLR044C	PDC1	● ■ ▲	● ■ ▲
	YLR134W	PDC5	○ □ ▲	● □ ▲
PFK	YMR205C	PFK2	● ■ ▲	● ■ ▲
	COMPLEX I	COMPLEX I	● □ △	● □ △
PGL	YGR248W	SOL4	○ □ △	● □ ▲
	YHR163W	SOL3	● ■ ▲	● ■ ▲
PGM1	YMR105C	PGM2	● □ ▲	● □ △
	YKL127W	PGM1	● □ △	● □ △
PYC	YBR218C	PYC2	● □ △	● □ △
	YGL062W	PYC1	● ■ ▲	● ■ ▲
SDH	COMPLEX I	COMPLEX I	● □ ▲	● □ △
	COMPLEX II	COMPLEX II	● ■ ▲	● ■ ▲
TKLa	YBR117C	TKL2	○ □ ▲	● □ △
	YPR074C	TKL1	● ■ △	● ■ ▲
TKLb	YBR117C	TKL2	○ □ ▲	● □ △
	YPR074C	TKL1	● ■ △	● ■ ▲

○ Proteomics
□ ecModel
△ Hybrid model

Absence Presence

Figure 31.- Comparison of protein allocation predictions by the hybrid and metabolic model in respiratory and mixed respire-fermentative metabolism. A) Boxplots for the distributions of log_{10} ratio between predicted and experimental abundance for enzymes in the enzyme-constrained model module, with and without the regulation layer. Respiration conditions

correspond to cells growing in chemostats at $0.1h^{-1}$ dilution rate, fermentation condition corresponds to chemostat cultures at $0.4h^{-1}$ dilution rate. Whiskers indicate the position of the lower and upper quartiles; the boxes, the distribution of the data points within the interquartile region; and horizontal lines indicate the position of the median for each distribution. B) Evaluation of utilization of isoenzymes, comparing the pure enzyme-constrained and the hybrid model vs. experimental data on protein expression (presence/absence). FMI.- Fowlkes-Mallows index. C) Comparison of predicted enzyme usages and protein expression data for each of the proteins in the pure enzyme-constrained and hybrid models. Black color indicates presence of a given protein.

6.3 The impact of glucose signaling over metabolic flux

The flux distributions used for comparison of protein predictions were mapped into network representations of the studied metabolic pathways. Reaction fold-changes between the predictions of the hybrid model and the pure enzyme-constrained model were computed in order to understand the points in metabolism that are affected by the regulatory and transcriptional layers. Larger regulatory effects were obtained in the low-glucose, respiratory condition, with an average flux fold-change of 1.85. A value of 0.46 was obtained for fermentation and high glucose availability condition, in contrast. This general trend reflects an overall effect of metabolic flux upregulation in conditions of low glucose, in comparison to an optimal flux distribution, whilst a general flux downregulation is induced under conditions of high glucose availability. This result was majorly caused by the number of reaction fluxes that are totally activated or deactivated by the regulatory layer (57 in low glucose and respiration, and 29 for high glucose fermentative conditions).

Figure 32 shows that the most notable effects of regulation over metabolic flux in respiration, were found around the pyruvate node. Interestingly, this metabolite is the intersection of multiple pathways, being the end product of glycolysis, from which the carbon flux can go either into fermentative pathway producing ethanol and acetate, or towards the TCA cycle, for cellular respiration and production of metabolic building blocks. In parallel, the flux through the phosphoenolpyruvate carboxykinase, which regenerates phosphoenolpyruvate from oxaloacetate, a crucial reaction for gluconeogenesis, was found to be highly upregulated. These results suggest that under scarcity of glucose, *S. cerevisiae* activates reaction steps that are fundamental for metabolic switches that may require regeneration of glycolytic intermediates from non-fermentable carbon sources.

Additionally, as the formalism used in the transcriptional layer of the hybrid model (equations 6.1 and 6.2) may force the use of enzymes that are not required to sustain an optimal flux distribution, an excess of enzyme mass is predicted to be available for several reactions in the network. This effect induces the emergence of futile fluxes (activation of some reaction steps in both the backwards and forwards reaction) across the network. In respiratory conditions, this is clearly observed in galactose metabolism, as all its constituent enzymes are expressed, resulting in a net flux of zero mmol. Other large futile fluxes arise in the two final steps of TCA cycle,

reactions in lower glycolysis, and in alcohol dehydrogenase (ADH). These futile fluxes indicate that the regulatory layer induces expression of excess enzyme mass in reaction steps that ensure the flux of carbon all the way down through glycolysis in combination with a clockwise functioning of the TCA cycle, which is known as cellular respiration, being the most efficient pathway for substrate utilization for energy production.

Moreover, the futile fluxes through the ADH reaction and in galactose metabolism suggest that, under glucose limitations, the regulatory machinery of the cell ensures expression of the necessary enzymes for catabolizing other sugars or non-fermentable carbon sources. Interestingly, futile fluxes that in TDH, PGK and ENO, together with upregulation of PCK could allow to run flux towards gluconeogenesis, additionally additional enzyme expression in the TCA cycle favors their reversibility . These finding are also supported by expression data. Altogether, this behavior shows that *S. cerevisiae* cells have developed multi-responsive regulatory mechanisms that control flux through respiration, while expressing an enzymatic machinery that is ready to respond to changes in extracellular nutrients.

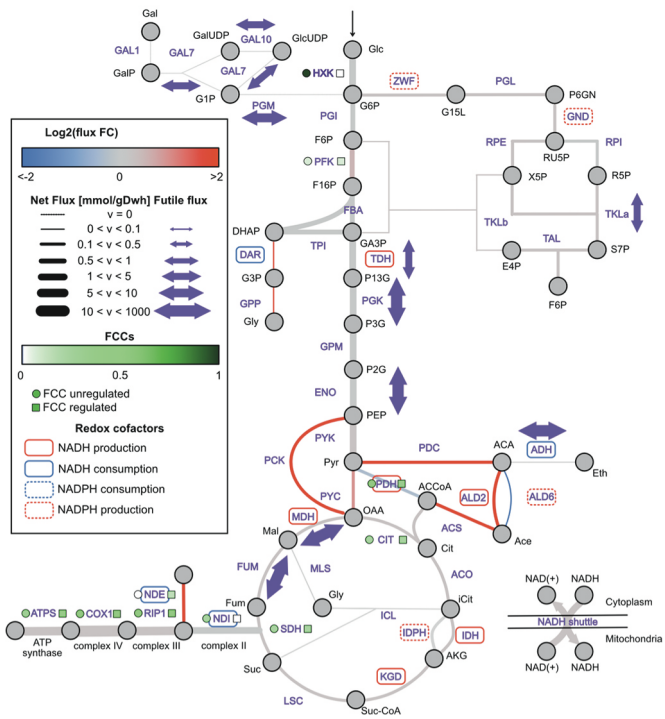


Figure 32.- Computing the impact of regulatory layer in flux predictions of the hybrid model under respiratory and low glucose conditions. Nodes indicate metabolites and edges the reactions catalyzed by the enzymes. Edge thickness indicates the magnitude of net flux values, color indicates the fold-change in predicted flux the hybrid model over the one from the pure enzyme-constrained model. Purple arrows show predicted futile fluxes. Green circles indicate

enzymes that exert control over the overall glucose uptake in the pure enzyme-constrained model. Green squares indicate enzymes that exert control over the overall glucose uptake in the hybrid model.

In respire-fermentative conditions at a high-glucose regime, an overall flux downregulation of the TCA cycle and oxidative phosphorylation pathways, together with upregulation of the fermentative fluxes were predicted by the hybrid model. Less futile fluxes are predicted to arise in this condition, nonetheless, several futile fluxes are predicted in glycolysis, with the highest one arising in the triose-phosphate isomerase reaction, known for its low equilibrium constant and its operation under near-to-equilibrium conditions[87,221,222]. Operation of biochemical reactions near to equilibrium creates conditions in which slight changes in intracellular metabolites can induce drastic changes in flux or even reverse the direction of the net flux, providing cells with metabolic robustness that can respond to environmental changes without the need of modulating translation rates of these enzymes[14,222]. Interestingly, the pattern of futile fluxes emerging due to regulation in respiratory metabolism, indicates that the cell is capable of adapting to galactose consumption, if needed, by expression of the Leloir pathway.

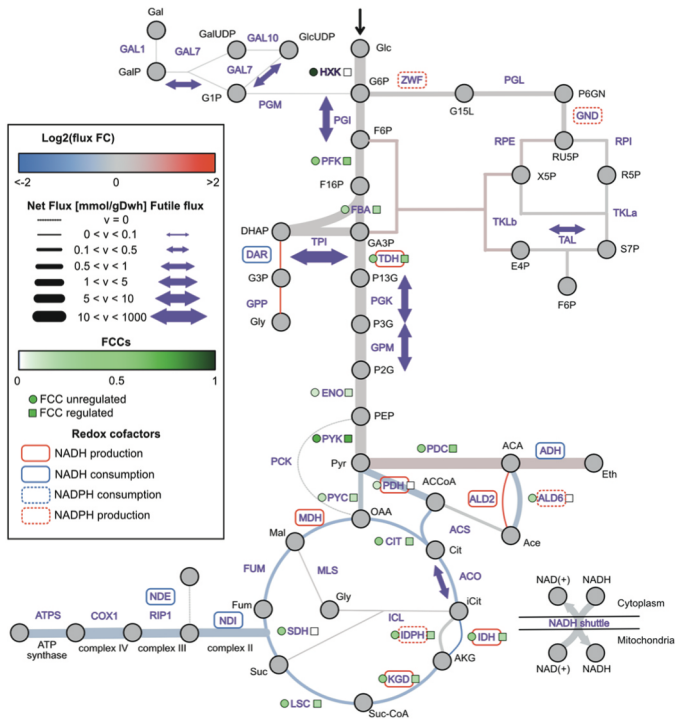


Figure 33.- Computing the impact of regulatory layer in flux predictions of the hybrid model under mixed respire-fermentative metabolism and high glucose conditions.

In order to analyze the effect of enzyme levels over the control of metabolic flux, flux control coefficients, defined by MCA theory, were approximated by running sensitivity analysis of glucose uptake rate to small perturbations on individual enzyme activities under steady state. When the approximated control coefficients are calculated with the pure enzyme-constrained model, a very high value (close to 1) is obtained for the hexokinase reaction, despite the glucose availability level. In contrast, computation of these coefficients with the hybrid model predicts a null control exerted by hexokinase, therefore, glucose uptake control is spread across more enzymes in the network. In low-glucose conditions control is mostly governed by enzymes in oxidative phosphorylation. In high-glucose respiro-fermentative conditions, control is spread across reaction steps that connect different pathways (PFK, FBA and TDH, connecting glycolysis and the PPP; and PYK, PDC and PYC around the pyruvate node, which connects three different pathways).

6.4 Summary

In **paper VI** a hybrid model, representing the regulatory gene network induced by nutrient signaling and an enzyme-constrained network of central carbon and energy metabolism in S. cerevisiae cells was constructed. The model was used to investigate the effects of gene regulation over metabolic flux.

Comparison of predictions between the pure enzyme-constrained model and the hybrid model revealed significant improvements in prediction of protein allocation to metabolic enzymes. This was mostly caused by diversified use of isoenzymes for multiple reaction steps in the hybrid model, in contrast to a pure enzyme-constrained model which predicts an optimal allocation of the protein and carbon resources. This finding is consistent with protein abundance data for respiratory and respire-fermentative conditions in which multiple isoforms are found to be expressed simultaneously.

The emergence of futile fluxes in different sectors of metabolism, especially under respiratory conditions, suggest that the regulation machinery might be able to leverage reactions operating near to equilibrium, in order to reduce the need for rapidly changing protein levels when nutrients are scarce. In contrast, the prediction of less futile fluxes and flux enzyme control spread across different pathways, predicted by the hybrid model under high glucose conditions, suggests that the regulatory machinery of yeast prioritizes metabolic control by modulating expression of key enzymes in situations of high protein demands.

Overall, construction of a structured model of gene regulation enables connection of expression of individual enzymes to specific steps in the regulatory network, offering a tool for a more comprehensive study of allocation of limited resources in the cell.

Conclusions

In this thesis I have explored how proteins constrain the metabolism and functioning of three different species of budding yeasts, *S. cerevisiae, Y. lipolytica* and *K. marxianus*. Different quantitative methods in systems biology were developed, extended, and implemented for generating an understanding of these constraints across different levels or scales of biological phenomena, such as evolution, metabolism, and regulation of protein expression. Moreover, a personal historical account of the evolution of systems biology and its core ideas and methods, was used as a mean to understand the possibilities, limitations, and implications of top-down and bottom-up approaches, both being used in this work.

The findings in **paper I** offered an evolutionary perspective of the genome and proteome of budding yeasts. Top-down analysis of their transcriptome and proteome under different conditions of environmental stress, indicated that exploration of the sequence space, by differentiated expression of small and non-essential proteins has enabled their adaptation to diverse ecological niches.

Top-down analyses can be used for understanding the presence and levels of gene products in cells. Bottom-up methods, such as metabolic modeling and simulation offer a quantitative overview of cellular processes. Integration of kinetic and protein abundance data into metabolic models has been used in this thesis to refine phenotype predictions across yeast species. Study of the protein abundance data into a metabolic context confirmed that, despite their phylogenetic relation, the high conservation of metabolism, and the common mechanism of stress adaptation, enzymatic limitations for long-term stress adaptation are different across these yeasts. Additionally, the enzyme limitations in metabolism under stress found with ecGEMs and proteomics data for these yeasts, showed to not correspond to their transcriptional responses, suggesting potential gene engineering targets for modulation of flux patterns. This can be leveraged for construction of robust strains for biotechnological processes, in which changing or stressful environments for cells are common.

In **paper IV** an ecGEM of *S. cerevisiae* was used to accelerate the design of a metabolic engineering strategy that increased intracellular accumulation of heme by 70-fold in cell cultures. The learnings from this project facilitated the development of a method of prediction of optimal metabolic engineering strategies for increasing production of metabolites in cells, that can account for allocation of carbon flux and the limited protein machinery of the cell towards the desired goal. Study of predicted gene engineering targets for increasing production of 100 different chemicals in *S. cerevisiae* cells, revealed that the complexity of the rewiring strategy increases for production of metabolites that come from highly enzymatically constrained pathways. Demonstrating that the design of production strains should account for the optimal way of balancing metabolic precursors, cofactors and a limited catalytic machinery between cellular growth and the production pathways.

Integration of a gene regulatory network of nutrient signaling mechanisms, together with an enzyme-constrained model of central carbon metabolism of *S. cerevisiae*, was used to study how cells modulate their enzymatic landscape to respond to changing levels of glucose in the environment. It was found that *S. cerevisiae* displays regulatory mechanisms that induce protein expression patterns that deviate from kinetic optimality, in order to provide the cell with robustness to changes in nutrient availability. Study of flux distributions showed that robustness can be gained by expressing enzymes for utilization of alternative carbon sources; modulated expression of enzymes with high kinetic control over the flux of carbon; and expression of excess enzyme mass in order to operate key reactions near to their equilibrium, to minimize the need of changes in translation if the metabolome changes.

The work in this thesis provides an overview of enzyme constraints in metabolism of budding yeast cells that can be described in a simple way by an analogy. The metabolic network is formed of interconnected "channels" (reactions) were nutrients flow. Different "turbines" are available in the network for generating the necessary energy for the maintenance and control of the flow through it. Several control systems exist for regulating the flow of nutrients through the network, "valves" can be opened, closed, or modulated in order to regulate the amount of flow through different sectors of the network of channels. These control systems can "sense" changes in the surroundings of the network, which is the source of the materials that flow through it.

Drastic changes in the upcoming material from the exterior trigger a cascade of events that regulate the "valves" inside the network, so that the energy turbines keep on functioning, and the flow of material irrigates essential points in the network. New channels are added randomly in peripheral sectors of the network over time. The valves in the new channels can be operated in diverse ways, and some of these will provide the network with capabilities for enduring other external changes in its surroundings. The additions of new channels and new ways of regulating the valves in them may derive in a network that can fulfill more functions that it was originally designed for. Finally, the network of channels can be leveraged for reaching different operation goals, by rational modification of the channels and controlling valves in it.

The use of different quantitative methods of systems biology in this thesis facilitated the identification of dual causality in biological processes. Cells modulate their gene expression according to the physico-chemical and environmental constraints that they are exposed to. Metabolic flux patterns that ensure survival of the cells under these environments, emerge as a property of the whole system, encompassing the changing exterior of the cell, and the modulated coordination of gene products in its interior (phenotype). Over long periods of time, these environmental changes can also perturb the genotype of the cell, which induces changes that translate in different and new accessible phenotypes.

In this thesis I have tried to demonstrate that systems biology provides a scientific framework for gaining novel understanding of biological processes, but also methodological tools that can be used for guiding rational manipulation of living cells for purposes of human interest, such as sustainable production of chemicals in microbial cells.

The historical narrative in the initial chapter of this thesis has the objective of showing that the development of a scientific discipline is a non-linear path. Multiple ideas and methods came to the study of biological processes from diverse fields of science and engineering, accumulation of these incorporations resulted in a shift of understanding of causational chains in living systems. Additionally, this historical review has made me realize that systems biology is a discipline with its own methodological, epistemological, and ontological propositions. Furthermore, parallel development and advancements in molecular biology continue nourishing systems biology. The increase in the available knowledge of biological components can be weaved together by the integrative approaches of systems biology to refine our understanding of life. Altogether, this has expanded biology, from a purely descriptive science to a discipline that can also be predictive and used rationally for manipulation of nature, as has also happened in the history of physics and chemistry.

The focus on mentioning specific names of scientists, their conceptual contributions to the history of systems biology, and their achievements, shows that a scientific discipline is something that evolves with time, by accumulation and interconnection of the work of multiple individuals and groups. Thus, the definition and description of a scientific field cannot be accomplished a priori and depends on an historical evolution process.

Finally, I would like to mention that the products of the work in this thesis have not only been used for the projects described here, but also for the rest of publications that I have contributed to during my years as a doctoral student. Even more, the resources produced here, including metabolic models, simulation methods, software for metabolic modeling and data analysis, etc. provide an infrastructure for facilitating the scientific endeavors of other individuals and groups in biological sciences.

Future perspectives

Systems biology has proven to be a powerful framework for integration of all the accumulated information and measurements of individual biological components into representations of living systems. Nevertheless, elucidation of biological functions from data analysis, or integration of data into models, depends largely on the prior knowledge on the function of genes. This results in difficulties for understanding the role of genetic components in higher-level functions. Different quantitative techniques should be combined to identify uncharacterized gene products, that are observed to be associated with emergent properties of living cells.

The top-down methods used in this thesis offer an example of how data-driven approaches can be used for narrowing down the efforts of characterizing genes of unknown function. Results from this project, and the posterior integration of proteomics data into metabolic models of yeast, highlight the need for a change of perspective to lead gene characterization endeavors, from asking what is the function of a gene? To asking, what properties or biological functions arise by the presence of this gene and its products in the context of a living system?

Mathematical modeling of cellular function, especially at genome-scale, is limited not just by the lack of knowledge in gene function, but also by the availability of kinetic and other physico-chemical parameters of gene products of known function. The advent of machine learning methods and its rapid integration into biological sciences has enormous potential for refinement of quantitative models in systems biology. Recently, multiple implementations of machine learning methods for prediction of kinetic parameters of enzymes have been developed. Extension and improvement of these methods will enhance the predictive power of enzyme-constrained models across organisms and diverse environments. Moreover, this offers the possibility of parameterizing more complex mathematical models of cell function, such as those encompassing metabolic and gene and protein expression processes, or even aiding the parameterization of large-scale kinetic models based on ODEs.

On parallel, further development of high-throughput techniques for facilitating and accelerating the measurement of physico-chemical parameters of gene products, is crucial for producing additional data to refine existing models, validate model predictions, and further enhance the training of machine-learning models for improved estimation of parameters.

Together, machine learning and experimental characterization of kinetic parameters of enzymes will refine the prediction of metabolic engineering strategies that balance the use of nutrients and cellular resources for a desired objective. Rewiring of metabolism has the potential of improving cell factories for bioproduction, engineering living species for bioremediation or climate change adaptation, and for medical applications in metabolic human diseases.

Incorporation of regulatory constraints into metabolic and enzyme constrained models has been used to improve phenotype predictions and provide understanding of emergent biological phenomena. Nonetheless, this approach has limited potential, as the knowledge on gene regulatory networks is mostly at the qualitative level. This problem can be partially tackled by using these hybrid models for identification of the components of gene regulation with the highest impact over phenotype predictions. This information can be used for identifying crucial components in gene regulation that can be studied with more quantitative experimental techniques. Iterations in this process could result in construction of quantitative understanding of the interactions between metabolism and gene regulation.

Acknowledgements

Becoming a PhD represents a major commitment to society, as it is them who choose to continue funding science and education. My profound gratitude to the Mexican, Swedish, and European societies who have provided the trust and funds for my education formation as a professional scientist.

My PhD journey could not have been possible without my supervisor. Jens, I will always be thankful for your trust, when you decided to take a student with no previous clue about biology. During all my years at SysBio it was your support, guidance and extreme patience what kept me going on. I have learned from you, not just from your brilliant scientific mind, but probably much more from you as a human being. You became a mentor in life for me, Thank you for everything.

To my cosupervisor Verena, all your help with the biological aspects of my research and your help throughout the CHASSY project have contributed to my development and realization of my projects. Eduard, also my cosupervisor, thank you for all your great feedback and fruitful discussions during the modeling group meetings and our manuscripts. I have learned a lot from you as a computational biologist and as a teacher. To my examiner and head of division, Ivan M., thanks for all your kindness, pragmatic piece of advice.

I have been extremely fortunate to have crossed my path in SysBio with those of relentless professors, researchers and postdoctors, from whom I tried to learn the most I could. Special thanks to Tyler, Lucy, Hao W., Kate, Jonathan, Daniel, Yu, Hongzhong, Rosemary, Rui, Manish, Yun, Stefan, Boyang, Fariba, Florian, Aleksej, Mikael and Martin.

Most of my studies were funded by the CHASSY project, an endeavor that pushed me to learn while delivering, in which I got the pleasure to be guided and supervised by John Morrisey, and enjoyed so many good trips and discussions with an amazing and fun team of people from all around Europe.

I also had the chance to share and create together with a prolific group of metabolic modelers that have been members of SysBio. Benjamín, I have had the best scientific conversations with you, thanks for your friendship and lessons on FBA. Feiran, I appreciate to have been able to share a lab with you and learn from your determination, you are a leading researcher in this discipline now. Avlant, thank you for your patience and always challenging questions and discussions. Dimitra, thanks for being the colleague that always offered a humane talk in the corridors. Mihail, our hours-long conversations fed my appetite for understanding every time they took place, let's keep on! Le, you have been a patient, understanding and caring office mate, thank you for all this my friend!

To all my PhD student peers across the years Parizad, Promi, Raphaël, Gang, Carl, Johan B., Yating, Johan G., Christoph, Kristos, Filip B., Linnéa, Kiana, My,

Maurizio, Kamesh, Cecilia, Marta, Oliver, Max, Dany, Andrés, Cheewin, Simone, Andrea, Angelo, Peishun, Hao L., Rassool, and a long list of etceteras. Sorry if I am missing names here but my time in SysBio has been long. I will always remember the shared times, games, dinners, discussions, parties, etc. You all PhDs are what keeps spinning the wheel of scientific discovery.

A large and prestigious research group, such as a SysBio, is formed by great scientists. Nevertheless, it is the work of all the administrators and support team what provides us with the best conditions and environment to practice our science. Martina, Gunilla, Erica, Anne-Lise, Jenny, Charlotte, Josefin, be sure that this thesis (and many others) would have never come to completion without your help, but most importantly for the care that you all have shown to me as a person.

To Efraín and Alberto, my former flat mates in Gothenburg, your patience and loyal friendship kept me sane and grounded. To Margareta, Klas and Anna Karin, you became my caring family in Gothenburg, thank you for accepting and welcoming me, you helped me enormously to cope with my homesickness. To Grete, you have shown me that a successful and determined scientist can also enjoy life and take care of oneself.

To my brothers in life, Fayez and Jesus, "la Sociedad de las Carpas Suicidas" is still alive, and I'm sure that we'll keep on walking and creating together for the rest of our lives. To my mentor Felipe, you are a great influence in my thinking, from you I have learned that modeling is a craft, that modeling is beautiful, that modeling is powerful, and that modeling is entertaining. You planted the seed of the seek for knowledge in my mind, gracias amigo! To the memory of Uriel, you taught me that an engineer has a social role to fulfil, your commitment to social justice shaped my ethics as a professional scientist, thank you for everything and for having inspired so many generations of creative minds.

To Louise, my words are small in comparison with everything you've done for me. You were there in the darkest and the brightest hours, you supported me and brought me out of my dark storms whenever I sank. Your company and care will never be forgotten by this heart. I admire you and want to see where your next steps guide you.

Finally, Mamá, Papá, Hermano, sin ustedes nada, me debo a ustedes, a su amor y apoyo incondicional. Gracias por apoyar cada una de mis disparatadas ideas y proyectos. Gracias por hacer todo esto posible y no detener mis sueños, gracias por hacer de mi la persona que soy hoy. Los amo.

References

1. McGovern, P. E. *et al*. Fermented beverages of pre- and proto-historic China. *Proc. Natl. Acad. Sci. U. S. A.* **101**, (2004).
2. Liu, L. *et al*. The origins of specialized pottery and diverse alcohol fermentation techniques in Early Neolithic China. *Proc. Natl. Acad. Sci. U. S. A.* **116**, (2019).
3. Chambers, P. J. & Pretorius, I. S. Fermenting knowledge: the history of winemaking, science and yeast research. *EMBO Rep.* **11**, 914–920 (2010).
4. SGD project. No Title. *Saccharomyces Genome Database*
5. Hohmann, S. Nobel Yeast Research. *FEMS Yeast Res.* **16**, fow094 (2016).
6. Nielsen, J. Yeast Systems Biology: Model Organism and Cell Factory. *Biotechnol. J.* **14**, 1800421 (2019).
7. Nonklang, S. *et al*. High-temperature ethanol fermentation and transformation with linear DNA in the thermotolerant yeast Kluyveromyces marxianus DMKU3-1042. *Appl. Environ. Microbiol.* (2008). doi:10.1128/AEM.01854-08
8. Palma, M. *et al*. The Zygosaccharomyces bailii transcription factor Haa1 is required for acetic acid and copper stress responses suggesting subfunctionalization of the ancestral bifunctional protein Haa1/Cup2. *BMC Genomics* (2017). doi:10.1186/s12864-016-3443-2
9. Nielsen, J. & Villadsen, J. *Bioreaction engineering principles Third Edition. Reactions* (2011).
10. Palsson, B. *Systems biology: Properties of reconstructed networks. Systems Biology: Properties of Reconstructed Networks* (2006). doi:10.1017/CBO9780511790515
11. Nielsen, J. Systems Biology of Metabolism. *Annu. Rev. Biochem.* **86**, 245–275 (2017).
12. Margulis, L., Sagan, C. & Sagan, D. Life. *Encyclopedia Britannica* https://www.britannica.com/science/life. (2023).
13. Teusink, B., Walsh, M. C., Van Dam, K. & Westerhoff, H. V. The danger of metabolic pathways with turbo design. *Trends Biochem. Sci.* (1998). doi:10.1016/S0968-0004(98)01205-5
14. Bar-Even, A., Flamholz, A., Noor, E. & Milo, R. Rethinking glycolysis: On the biochemical logic of metabolic pathways. *Nature Chemical Biology* **8**, (2012).
15. Lane, N. *Transformer: The Deep Chemistry of Life and Death*. (Profile, 2022).
16. Nelson 1942-, D. L. (David L. *Lehninger principles of biochemistry*. (Fourth edition. New York : W.H. Freeman, 2005., 2005).
17. Lane, N. The vital question: energy, evolution, and the origins of complex life. *Choice Rev. Online* **53**, (2016).
18. Saraste, M. Oxidative Phosphorylation at the fin de siècle. *Science (80-.).* **283**, 1488–1493 (1999).
19. Lane, N. Life force: why energy shapes evolution. *Biochem. (Lond).* **37**, 6–11 (2015).

20. Boyer, P. D. A research journey with ATP synthase. *Journal of Biological Chemistry* **277**, (2002).
21. Schrödinger, E. *What is Life? The Physical Aspect of the Living Cell*. (Cambridge University Press, 1944).
22. David, S. V. & Hayden, B. Y. Neurotree: A Collaborative, Graphical Database of the Academic Genealogy of Neuroscience. *PLoS One* **7**, (2012).
23. WATSON, J. D. & CRICK, F. H. C. Molecular Structure of Nucleic Acids: A Structure for Deoxyribose Nucleic Acid. *Nature* **171**, 737–738 (1953).
24. Strevens, M. *The knowledge machine: how irrationality created modern science*. (Liveright Publishing Corporation, 2020).
25. Crick, F. H. C. On Protein Syntesis. *Symp. Soc. Exp. Biol. XII* (1958).
26. Crick, F. Central dogma of molecular biology. *Nature* **227**, (1970).
27. Díez, J. Falsificationism and the structure of theories: the Popper-Kuhn controversy about the rationality of normal science. *Stud. Hist. Philos. Sci. Part A* **38**, (2007).
28. Morange, M. History of Molecular Biology. in *eLS* (2016). doi:10.1002/9780470015902.a0003079.pub3
29. Boogerd, F. C., Bruggeman, F. J., Hofmeyr, J.-H. S. & Westerhoff, H. V. *Systems biology: philosophical foundations*. *Systems Biology* (2007).
30. Kuhn, T. S. *The Structure of Scientific Revolutions*. **2**, (University of Chicago Press, 1962).
31. HUTTER, O. F. & NOBLE, D. Rectifying Properties of Heart Muscle. *Nature* **188**, 495 (1960).
32. NOBLE, D. Cardiac Action and Pacemaker Potentials based on the Hodgkin-Huxley Equations. *Nature* **188**, 495–497 (1960).
33. von Bertalanffy, L. The Theory of Open Systems in Physics and Biology. *Science (80-.).* **111**, 23–29 (1950).
34. Mitchell, P. Possible molecular mechanisms of the protonmotive function of cytochrome systems. *J. Theor. Biol.* **62**, 327–367 (1976).
35. Kell, D. B. On the functional proton current pathway of electron transport phosphorylation: an electrodic view. *Biochim. Biophys. Acta (BBA)-Reviews Bioenerg.* **549**, 55–99 (1979).
36. Van Dam, K. & Westerhoff, H. V. A description of oxidative phosphorylation in terms of irreversible thermodynamics. in *Structure and Function of Energy-transducing Membranes: Proceedings of a Workshop Held in Amsterdam on August 10-13, 1977, in Honour of EC Slater's 60th Birthday* **14**, 157 (Elsevier Science & Technology, 1977).
37. Heinrich, R. & Rapoport, T. A. A Linear Steady-State Treatment of Enzymatic Chains. *Eur. J. Biochem.* **42**, 89–95 (1974).
38. Rapoport, T. A., Höhne, W. E., Reich, J. G., Heitmann, P. & Rapoport, S. M. A Kinetic Model for the Action of the Inorganic Pyrophosphatase from Bakers' Yeast: The Activating Influence of Magnesium Ions. *Eur. J. Biochem.* **26**, 237–246 (1972).
39. Rapoport, T. A. & Heinrich, R. Mathematical analysis of multienzyme systems. I. Modelling of the glycolysis of human erythrocytes. *Biosystems*

7, 120–129 (1975).
40. Garfinkel, D., Garfinkel, L., Pring, M., Green, S. B. & Chance, B. Computer Applications to Biochemical Kinetics. *Annu. Rev. Biochem.* **39**, 473–498 (1970).
41. Kacser, H. & Burns, J. A. Rate control of biological processes. in *Symp. Soc. Exp. Biol* **27**, 65–104 (1973).
42. Brenner, S. ON THE IMPOSSIBILITY OF ALL OVERLAPPING TRIPLET CODES IN INFORMATION TRANSFER FROM NUCLEIC ACID TO PROTEINS. *Proc. Natl. Acad. Sci.* **43**, 687–694 (1957).
43. Wiener, N. Perspectives in Cybernetics. in (eds. Wiener, N. & Schadé, J. P. B. T.-P. in B. R.) **17**, 399–415 (Elsevier, 1965).
44. Chang, M. A., Dayhoff, M. O., Eck, R. V & Sochard, M. R. *Atlas of protein sequence and structure*. (1965).
45. Westerhoff, H. V. & Palsson, B. O. The evolution of molecular biology into systems biology. *Nature Biotechnology* **22**, 1249–1252 (2004).
46. Beadle, G. W. & Tatum, E. L. Genetic control of biochemical reactions in Neurospora. *Proc. Natl. Acad. Sci.* **27**, 499–506 (1941).
47. Kacser, H., Burns, J. A. & Fell, D. A. The control of flux. in *Biochemical Society Transactions* (1995). doi:10.1042/bst0230341
48. Maxam, A. M. & Gilbert, W. A new method for sequencing DNA. *Proc. Natl. Acad. Sci.* **74**, 560–564 (1977).
49. Sanger, F. & Coulson, A. R. A rapid method for determining sequences in DNA by primed synthesis with DNA polymerase. *J. Mol. Biol.* **94**, 441–448 (1975).
50. Lewis, S. E. Gene Ontology: looking backwards and forwards. *Genome Biol.* **6**, 103 (2004).
51. Fodor, S. P. A. *et al*. Multiplexed biochemical assays with biological chips. *Nature* **364**, 555–556 (1993).
52. Schena, M., Shalon, D., Davis, R. W. & Brown, P. O. Quantitative monitoring of gene expression patterns with a complementary DNA microarray. *Science (80-.).* **270**, 467–470 (1995).
53. Ashburner, M. *et al*. Gene ontology: Tool for the unification of biology. *Nat. Genet.* **25**, 25–29 (2000).
54. Gardner, T. S., Cantor, C. R. & Collins, J. J. Construction of a genetic toggle switch in Escherichia coli. *Nature* **403**, 339–342 (2000).
55. Becskei, A. & Serrano, L. Engineering stability in gene networks by autoregulation. *Nature* **405**, 590–593 (2000).
56. Uetz, P. *et al*. A comprehensive analysis of protein–protein interactions in Saccharomyces cerevisiae. *Nature* **403**, 623–627 (2000).
57. Jeong, H., Tombor, B., Albert, R., Oltvai, Z. N. & Barabási, A.-L. The large-scale organization of metabolic networks. *Nature* **407**, 651–654 (2000).
58. Walhout, A. J. M. *et al*. Protein interaction mapping in C. elegans using proteins involved in vulval development. *Science (80-.).* **287**, 116–122 (2000).
59. Ferrell Jr, J. E. & Machleder, E. M. The biochemical basis of an all-or-

none cell fate switch in Xenopus oocytes. *Science (80-.).* **280**, 895–898 (1998).
60. Ideker, T. *et al*. Integrated genomic and proteomic analyses of a systematically perturbed metabolic network. *Science (80-.).* **292**, 929–934 (2001).
61. Kitano, H. Perspectives on systems biology. *New Gener. Comput.* **18**, 199–216 (2000).
62. Hiroaki, K. *Foundations of systems biology*. (The MIT Press, 2001).
63. Ideker, T., Galitski, T. & Hood, L. A NEW APPROACH TO DECODING LIFE: Systems Biology. *Annu. Rev. Genomics Hum. Genet.* **2**, 343–372 (2001).
64. Kirschner, M. W. The meaning of systems biology. *Cell* **121**, 503–504 (2005).
65. Noble, D. *The music of life: biology beyond genes*. (Oxford University Press, USA, 2008).
66. Noble, D. Claude Bernard, the first systems biologist, and the future of physiology. *Exp. Physiol.* **93**, 16–26 (2008).
67. Noble, D. A theory of biological relativity: no privileged level of causation. *Interface Focus* **2**, 55–64 (2012).
68. Palsson, B. Ø. *Systems Biology: Constraint-based Reconstruction and Analysis*. (Cambridge University Press, 2015). doi:DOI: 10.1017/CBO9781139854610
69. Fleischmann, R. D. *et al*. Whole-genome random sequencing and assembly of Haemophilus influenzae Rd. *Science (80-.).* (1995). doi:10.1126/science.7542800
70. Goffeau, A. *et al*. Life with 6000 genes. *Science (80-.).* **274**, 546–567 (1996).
71. Monk, J. M. *et al*. Multi-omics Quantification of Species Variation of Escherichia coli Links Molecular Features with Strain Phenotypes. *Cell Syst.* **3**, 238-251.e12 (2016).
72. Blattner, F. R. *et al*. The complete genome sequence of Escherichia coli K-12. *Science (80-.).* **277**, 1453–1462 (1997).
73. Goodwin, S., McPherson, J. D. & McCombie, W. R. Coming of age: ten years of next-generation sequencing technologies. *Nat. Rev. Genet.* **17**, 333–351 (2016).
74. Levy, S. E. & Myers, R. M. Advancements in Next-Generation Sequencing. *Annu. Rev. Genomics Hum. Genet.* **17**, 95–115 (2016).
75. Shapiro, E., Biezuner, T. & Linnarsson, S. Single-cell sequencing-based technologies will revolutionize whole-organism science. *Nat. Rev. Genet.* **14**, 618–630 (2013).
76. Navin, N. *et al*. Tumour evolution inferred by single-cell sequencing. *Nature* **472**, 90–94 (2011).
77. Millán-Oropeza, A., Blein-Nicolas, M., Monnet, V., Zivy, M. & Henry, C. Comparison of Different Label-Free Techniques for the Semi-Absolute Quantification of Protein Abundance. *Proteomes* **10**, (2022).
78. Sánchez, B. J. *et al*. Benchmarking accuracy and precision of intensity-

based absolute quantification of protein abundances in Saccharomyces cerevisiae. *Proteomics* **21**, (2021).
79. O'Brien, E. J., Utrilla, J. & Palsson, B. O. Quantification and Classification of E. coli Proteome Utilization and Unused Protein Costs across Environments. *PLoS Comput. Biol.* **12**, (2016).
80. Di Bartolomeo, F. *et al.* Absolute yeast mitochondrial proteome quantification reveals trade-off between biosynthesis and energy generation during diauxic shift. *Proc. Natl. Acad. Sci. U. S. A.* (2020). doi:10.1073/pnas.1918216117
81. Arike, L. *et al.* Comparison and applications of label-free absolute proteome quantification methods on Escherichia coli. *J. Proteomics* (2012). doi:10.1016/j.jprot.2012.06.020
82. Wiśniewski, J. R., Vildhede, A., Norén, A. & Artursson, P. In-depth quantitative analysis and comparison of the human hepatocyte and hepatoma cell line HepG2 proteomes. *J. Proteomics* (2016). doi:10.1016/j.jprot.2016.01.016
83. Björkeroth, J. *et al.* Proteome reallocation from amino acid biosynthesis to ribosomes enables yeast to grow faster in rich media. *Proc. Natl. Acad. Sci.* **117**, 21804 LP – 21812 (2020).
84. Fuhrer, T. & Zamboni, N. High-throughput discovery metabolomics. *Curr. Opin. Biotechnol.* **31**, 73–78 (2015).
85. Zampieri, M., Sekar, K., Zamboni, N. & Sauer, U. Frontiers of high-throughput metabolomics. *Curr. Opin. Chem. Biol.* **36**, 15–23 (2017).
86. Nielsen, J. It Is All about Metabolic Fluxes. *Journal of Bacteriology* **185**, 7031–7035 (2003).
87. Xu, J. *et al.* Metabolic flux analysis and fluxomics-driven determination of reaction free energy using multiple isotopes. *Current Opinion in Biotechnology* (2020). doi:10.1016/j.copbio.2020.02.018
88. Heux, S., Bergès, C., Millard, P., Portais, J.-C. & Létisse, F. Recent advances in high-throughput 13C-fluxomics. *Curr. Opin. Biotechnol.* **43**, 104–109 (2017).
89. Niedenführ, S., Wiechert, W. & Nöh, K. How to measure metabolic fluxes: a taxonomic guide for 13C fluxomics. *Curr. Opin. Biotechnol.* **34**, 82–90 (2015).
90. Cammack, R. *et al.* +ome. (2008). doi:10.1093/acref/9780198529170.013.14194
91. Bruggeman, F. J. & Westerhoff, H. V. The nature of systems biology. *TRENDS Microbiol.* **15**, 45–50 (2007).
92. Nielsen, J. & Jewett, M. C. Impact of systems biology on metabolic engineering of Saccharomyces cerevisiae. *FEMS Yeast Research* **8**, 122–131 (2008).
93. Doughty, T. W. *et al.* Stress-induced expression is enriched for evolutionarily young genes in diverse budding yeasts. *Nat. Commun.* (2020). doi:10.1038/s41467-020-16073-3
94. Alonso-Gutierrez, J. *et al.* Principal component analysis of proteomics (PCAP) as a tool to direct metabolic engineering. *Metab. Eng.* **28**, 123–133

(2015).
95. Väremo, L., Nielsen, J. & Nookaew, I. Enriching the gene set analysis of genome-wide data by incorporating directionality of gene expression and combining statistical hypotheses and methods. *Nucleic Acids Res*. **41**, 4378–4391 (2013).
96. Lahtvee, P. J. *et al*. Absolute Quantification of Protein and mRNA Abundances Demonstrate Variability in Gene-Specific Translation Efficiency in Yeast. *Cell Syst*. **4**, 495-504.e5 (2017).
97. Ebrahim, A. *et al*. Multi-omic data integration enables discovery of hidden biological regularities. *Nat*. *Commun*. **7**, (2016).
98. Xia, J. *et al*. Proteome allocations change linearly with the specific growth rate of Saccharomyces cerevisiae under glucose limitation. *Nat*. *Commun*. **13**, 2819 (2022).
99. Campbell, K. *et al*. Building blocks are synthesized on demand during the yeast cell cycle. *Proc*. *Natl*. *Acad*. *Sci*. *U*. *S*. *A*. (2020). doi:10.1073/pnas.1919535117
100. Vidal, M. A unifying view of 21st century systems biology. *FEBS Letters* **583**, 3891–3894 (2009).
101. Alberich, R., Castro, J. A., Llabres, M. & Palmer-Rodriguez, P. Metabolomics analysis: Finding out metabolic building blocks. *PLoS One* **12**, e0177031 (2017).
102. Kotze, H. L. *et al*. A novel untargeted metabolomics correlation-based network analysis incorporating human metabolic reconstructions. *BMC Syst*. *Biol*. **7**, 1–11 (2013).
103. Frainay, C. & Jourdan, F. Computational methods to identify metabolic sub-networks based on metabolomic profiles. *Brief*. *Bioinform*. **18**, 43–56 (2017).
104. Shen, Z., Bao, W. & Huang, D.-S. Recurrent neural network for predicting transcription factor binding sites. *Sci*. *Rep*. **8**, 15270 (2018).
105. Bergenholm, D., Liu, G., Holland, P. & Nielsen, J. Reconstruction of a global transcriptional regulatory network for control of lipid metabolism in yeast by using chromatin immunoprecipitation with lambda exonuclease digestion. *Msystems* **3**, e00215-17 (2018).
106. Kell, D. B. & Oliver, S. G. Here is the evidence, now what is the hypothesis? The complementary roles of inductive and hypothesis-driven science in the post-genomic era. *Bioessays* **26**, 99–105 (2004).
107. Novick, A. & Weiner, M. Enzyme induction as an all-or-none phenomenon. *Proc*. *Natl*. *Acad*. *Sci*. **43**, 553–566 (1957).
108. Von Bertalanffy, L. Quantitative laws in metabolism and growth. *Q*. *Rev*. *Biol*. **32**, 217–231 (1957).
109. Tsuchiya, H. M., Fredrickson, A. G. & Aris, R. Dynamics of microbial cell populations. in *Advances in chemical engineering* **6**, 125–206 (Elsevier, 1966).
110. Fredrickson, A. G., Ramkrishna, D. & Tsuchiya, H. M. Statistics and dynamics of procaryotic cell populations. *Math*. *Biosci*. **1**, 327–374 (1967).
111. Jost, J. L., Drake, J. F., Fredrickson, A. G. & Tsuchiya, H. M. Interactions

of Tetrahymena pyriformis, Escherichia coli, Azotobacter vinelandii, and glucose in a minimal medium. *J. Bacteriol.* **113**, 834–840 (1973).
112. Stephanopoulos, G., Aris, R. & Fredrickson, A. G. A stochastic analysis of the growth of competing microbial populations in a continuous biochemical reactor. *Math. Biosci.* **45**, 99–135 (1979).
113. Westerhoff, H. V, van Heeswijk, W., Kahn, D. & Kell, D. B. Quantitative approaches to the analysis of the control and regulation of microbial metabolism. *Quant. Asp. Growth Metab. Microorg.* 193–207 (1992).
114. Pir, P. *et al.* The genetic control of growth rate: a systems biology study in yeast. *BMC Syst. Biol.* **6**, 1–17 (2012).
115. Delneri, D. *et al.* Identification and characterization of high-flux-control genes of yeast through competition analyses in continuous cultures. *Nat. Genet.* **40**, 113–117 (2008).
116. Kell, D. B. & Westerhoff, H. V. Metabolic control theory: its role in microbiology and biotechnology. *FEMS Microbiol. Rev.* **2**, 305–320 (1986).
117. Aris, R. & Gavalas, G. R. On the theory of reactions in continuous mixtures. *Philos. Trans. R. Soc. London. Ser. A, Math. Phys. Sci.* **260**, 351–393 (1966).
118. Wei, J. & Prater, C. D. The structure and analysis of complex reaction systems. in *Advances in catalysis* **13**, 203–392 (Elsevier, 1962).
119. Papoutsakis, E. T. Equations and calculations for fermentations of butyric acid bacteria. *Biotechnol. Bioeng.* **26**, (1984).
120. Watson, M. R. Metabolic maps for the Apple II. (1984).
121. Palsson, B. O. & Lightfoot, E. N. Mathematical modelling of dynamics and control in metabolic networks. I. On michaelis-menten kinetics. *J. Theor. Biol.* **111**, (1984).
122. Joshi, A. & Palsson, B. O. Metabolic dynamics in the human red cell: Part I—A comprehensive kinetic model. *J. Theor. Biol.* **141**, 515–528 (1989).
123. Varma, A. & Palsson, B. O. Metabolic capabilities of escherichia coli. II. Optimal growth patterns. *J. Theor. Biol.* **165**, (1993).
124. Varma, A. & Palsson, B. O. Stoichiometric flux balance models quantitatively predict growth and metabolic by-product secretion in wild-type Escherichia coli W3110. *Appl. Environ. Microbiol.* **60**, 3724–3731 (1994).
125. Varma, A., Boesch, B. W. & Palsson, B. O. Stoichiometric interpretation of Escherichia coli glucose catabolism under various oxygenation rates. *Appl. Environ. Microbiol.* **59**, 2465–2473 (1993).
126. Varma, A. & Palsson, B. O. Metabolic capabilities of Escherichia coli: I. Synthesis of biosynthetic precursors and cofactors. *J. Theor. Biol.* **165**, 477–502 (1993).
127. Edwards, J. S. & Palsson, B. O. Systems properties of the Haemophilus influenzaeRd metabolic genotype. *J. Biol. Chem.* **274**, 17410–17416 (1999).
128. Heinrich, R. & Klipp, E. Control analysis of unbranched enzymatic chains in states of maximal activity. *J. Theor. Biol.* **182**, 243–252 (1996).

129. KLIPP, E. EVOLUTIONARY OPTIMIZATION OF ENZYME KINETIC PARAMETERS. *J. Biol. Syst.* **03**, (1995).
130. Hatzimanikatis, V. & Bailey, J. E. Effects of spatiotemporal variations on metabolic control: approximate analysis using (log) linear kinetic models. *Biotechnol. Bioeng.* **54**, 91–104 (1997).
131. Hatzimanikatis, V. Nonlinear Metabolic Control Analysis. *Metab. Eng.* **1**, 75–87 (1999).
132. Hatzimanikatis, V. & Bailey, J. E. Studies on glycolysis—I. Multiple steady states in bacterial glycolysis. *Chem. Eng. Sci.* **52**, 2579–2588 (1997).
133. Teusink, B., Bakker, B. M. & Westerhoff, H. V. Control of frequency and amplitudes is shared by all enzymes in three models for yeast glycolytic oscillations. *Biochim. Biophys. Acta (BBA)-Bioenergetics* **1275**, 204–212 (1996).
134. Richard, P., Teusink, B., Westerhoff, H. V & van Dam, K. Around the growth phase transition S. cerevisiae's make-up favours sustained oscillations of intracellular metabolites. *FEBS Lett.* **318**, 80–82 (1993).
135. Richard, P., Bakker, B. M., Teusink, B., Van Dam, K. & Westerhoff, H. V. Acetaldehyde mediates the synchronization of sustained glycolytic oscillations in populations of yeast cells. *Eur. J. Biochem.* **235**, 238–241 (1996).
136. Teusink, B., Baganz, F., Westerhoff, H. V & Oliver, S. G. 17 Metabolic Control Analysis as a Tool in the Elucidation of the Function of Novel Genes. in *Methods in microbiology* **26**, 297–336 (Elsevier, 1998).
137. Teusink, B. & Westerhoff, H. V. 'Slave'metabolites and enzymes: A rapid way of delineating metabolic control. *Eur. J. Biochem.* **267**, 1889–1893 (2000).
138. Shuler, M. L., Leung, S. & Dick, C. C. A mathematical model for the growth of a single bacterial cell. *Ann. N. Y. Acad. Sci.* **326**, 35–52 (1979).
139. Tomita, M. *et al.* E-CELL: software environment for whole-cell simulation. *Bioinformatics* **15**, 72–84 (1999).
140. Karr, J. R., Sanghvi, J. C., Jacobs, J. M., Macklin, D. N. & Covert, M. W. A whole cell model of mycoplasma genitalium elucidates mechanisms of bacterial replication. *Biophys. J.* **102**, 731a (2012).
141. Karr, J. R. *et al.* A whole-cell computational model predicts phenotype from genotype. *Cell* **150**, 389–401 (2012).
142. Ye, C. *et al.* Comprehensive understanding of Saccharomyces cerevisiae phenotypes with whole-cell model WM_S288C. *Biotechnol. Bioeng.* **117**, 1562–1574 (2020).
143. Thiele, I. & Palsson, B. O. *Introduction to Systems Biology. Introduction to Systems Biology* (2007). doi:10.1007/978-1-59745-531-2
144. Orth, J. D., Thiele, I. & Palsson, B. Ø. What is flux balance analysis? *Nat. Biotechnol.* **28**, 245–248 (2010).
145. Feist, A. M. & Palsson, B. O. The biomass objective function. *Current Opinion in Microbiology* **13**, (2010).
146. Lewis, N. E., Nagarajan, H. & Palsson, B. O. Constraining the metabolic

genotype-phenotype relationship using a phylogeny of in silico methods. *Nat. Rev. Microbiol.* **10**, 291–305 (2012).
147. Schellenberger, J. *et al.* Quantitative prediction of cellular metabolism with constraint-based models: the COBRA Toolbox v2.0. *Nat. Protoc.* **6**, 1290–1307 (2011).
148. Mahadevan, R. & Schilling, C. H. The effects of alternate optimal solutions in constraint-based genome-scale metabolic models. *Metab. Eng.* **5**, 264–276 (2003).
149. Beg, Q. K. *et al.* Intracellular crowding defines the mode and sequence of substrate uptake by Escherichia coli and constrains its metabolic activity. *Proc. Natl. Acad. Sci. U. S. A.* **104**, 12663–12668 (2007).
150. Vazquez, A. *et al.* Impact of the solvent capacity constraint on E. coli metabolism. *BMC Syst. Biol.* **2**, 7 (2008).
151. Zhuang, K., Vemuri, G. N. & Mahadevan, R. Economics of membrane occupancy and respiro-fermentation. *Mol. Syst. Biol.* **7**, (2011).
152. Adadi, R., Volkmer, B., Milo, R., Heinemann, M. & Shlomi, T. Prediction of microbial growth rate versus biomass yield by a metabolic network with kinetic parameters. *PLoS Comput. Biol.* **8**, (2012).
153. Shlomi, T., Benyamini, T., Gottlieb, E., Sharan, R. & Ruppin, E. Genome-scale metabolic modeling elucidates the role of proliferative adaptation in causing the warburg effect. *PLoS Comput. Biol.* **7**, 1–8 (2011).
154. Nilsson, A. & Nielsen, J. Metabolic Trade-offs in Yeast are Caused by F1F0-ATP synthase. *Sci. Rep.* **6**, 1–11 (2016).
155. Chen, Y. & Nielsen, J. Energy metabolism controls phenotypes by protein efficiency and allocation. *Proc. Natl. Acad. Sci. U. S. A.* (2019). doi:10.1073/pnas.1906569116
156. Sánchez, B. J. *et al.* Improving the phenotype predictions of a yeast genome-scale metabolic model by incorporating enzymatic constraints. *Mol. Syst. Biol.* **13**, 935 (2017).
157. Domenzain, I. *et al.* Reconstruction of a catalogue of genome-scale metabolic models with enzymatic constraints using GECKO 2.0. *Nat. Commun.* **13**, 3766 (2022).
158. Nilsson, A., Björnson, E., Flockhart, M., Larsen, F. J. & Nielsen, J. Complex I is bypassed during high intensity exercise. *Nat. Commun.* (2019). doi:10.1038/s41467-019-12934-8
159. Lu, H. *et al.* A consensus S. cerevisiae metabolic model Yeast8 and its ecosystem for comprehensively probing cellular metabolism. *Nat. Commun.* (2019). doi:10.1038/s41467-019-11581-3
160. Robinson, J. L. *et al.* An atlas of human metabolism. *Sci. Signal.* (2020). doi:10.1126/scisignal.aaz1482
161. Ma, Y.-J., Lin, L.-L., Chien, H. R. & Hsu, W.-H. Efficient utilization of starch by a recombinant strain of Saccharomyces cerevisiae producing glucoamylase and isoamylase. *Biotechnol. Appl. Biochem.* **31**, (2000).
162. Morrissey, J. P., Varela, J. A., Gethins, L., Stanton, C. & Ross, P. Applications of kluyveromyces marxianus in biotechnology. in *Yeast Diversity in Human Welfare* (2017). doi:10.1007/978-981-10-2621-8_17

163. Ledesma-Amaro, R. & Nicaud, J. M. Yarrowia lipolytica as a biotechnological chassis to produce usual and unusual fatty acids. *Progress in Lipid Research* **61**, (2016).
164. Deparis, Q., Claes, A., Foulquié-Moreno, M. R. & Thevelein, J. M. Engineering tolerance to industrially relevant stress factors in yeast cell factories. *FEMS yeast research* **17**, (2017).
165. Reyes-Rosales, A. *et al*. Identification of genetic and biochemical mechanisms associated with heat shock and heat stress adaptation in grain amaranths. *Front. Plant Sci.* **14**, 1101375 (2023).
166. Zybailov, B. *et al*. Statistical analysis of membrane proteome expression changes in Saccharomyces cerevisiae. *J. Proteome Res.* (2006). doi:10.1021/pr060161n
167. Howe, K. L. *et al*. Ensembl 2021. *Nucleic Acids Res.* **49**, D884–D891 (2021).
168. Conesa, A. & Götz, S. Blast2GO: a comprehensive suite for functional analysis in plant genomics. *Int. J. Plant Genomics* **2008**, (2008).
169. Shen, X. X. *et al*. Tempo and Mode of Genome Evolution in the Budding Yeast Subphylum. *Cell* (2018). doi:10.1016/j.cell.2018.10.023
170. Emms, D. M. & Kelly, S. OrthoFinder: phylogenetic orthology inference for comparative genomics. *Genome Biol.* **20**, 1–14 (2019).
171. Benjamini, Y. & Hochberg, Y. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *J. R. Stat. Soc. Ser. B* (2018). doi:10.1111/j.2517-6161.1995.tb02031.x
172. Pál, C., Papp, B. & Hurst, L. D. Highly expressed genes in yeast evolve slowly. *Genetics* **158**, 927–931 (2001).
173. Mata, J. & Bähler, J. Correlations between gene expression and gene conservation in fission yeast. *Genome Res.* **13**, 2686–2690 (2003).
174. Förster, J., Famili, I., Fu, P., Palsson, B. & Nielsen, J. Genome-scale reconstruction of the Saccharomyces cerevisiae metabolic network. *Genome Res.* (2003). doi:10.1101/gr.234503
175. Sánchez, B. J. & Nielsen, J. Genome scale models of yeast: towards standardized evaluation and consistent omic integration. *Integr. Biol. (United Kingdom)* (2015). doi:10.1039/c5ib00083a
176. Lopes, H. & Rocha, I. Genome-scale modeling of yeast: chronology, applications and critical perspectives. *FEMS yeast research* (2017). doi:10.1093/femsyr/fox050
177. Castillo, S., Patil, K. R. & Jouhten, P. Yeast Genome-Scale Metabolic Models for Simulating Genotype-Phenotype Relations. *Progress in molecular and subcellular biology* (2019). doi:10.1007/978-3-030-13035-0_5
178. Chen, Y., Li, G. & Nielsen, J. Genome-Scale Metabolic Modeling from Yeast to Human Cell Models of Complex Diseases: Latest Advances and Challenges. *Methods Mol. Biol.* **2049**, 329–345 (2019).
179. Herrgård, M. J. *et al*. A consensus yeast metabolic network reconstruction obtained from a community approach to systems biology. *Nature Biotechnology* **26**, 1155–1160 (2008).

180. Dobson, P. D. *et al*. Further developments towards a genome-scale metabolic model of yeast. *BMC Syst. Biol*. (2010). doi:10.1186/1752-0509-4-145
181. Heavner, B. D., Smallbone, K., Barker, B., Mendes, P. & Walker, L. P. Yeast 5 - an expanded reconstruction of the Saccharomyces cerevisiae metabolic network. *BMC Syst. Biol*. (2012). doi:10.1186/1752-0509-6-55
182. Heavner, B. D., Smallbone, K., Price, N. D. & Walker, L. P. Version 6 of the consensus yeast metabolic network refines biochemical coverage and improves model performance. *Database* (2013). doi:10.1093/database/bat059
183. Aung, H. W., Henry, S. A. & Walker, L. P. Revising the representation of fatty acid, glycerolipid, and glycerophospholipid metabolism in the consensus model of yeast metabolism. *Ind. Biotechnol*. (2013). doi:10.1089/ind.2013.0013
184. Chelliah, V. *et al*. BioModels: Ten-year anniversary. *Nucleic Acids Res*. (2015). doi:10.1093/nar/gku1181
185. Ebrahim, A., Lerman, J. A., Palsson, B. O. & Hyduke, D. R. COBRApy: COnstraints-Based Reconstruction and Analysis for Python. *BMC Syst. Biol*. (2013). doi:10.1186/1752-0509-7-74
186. Olivier, B. G. & Bergmann, F. T. SBML Level 3 Package: Flux Balance Constraints version 2 SBML Level 3 Package: Flux Balance Constraints ('fbc'). *J. Integr. Bioinform*. (2018). doi:10.1515/jib-2017-0082
187. Heirendt, L. *et al*. Creation and analysis of biochemical constraint-based models using the COBRA Toolbox v.3.0. *Nat. Protoc*. **14**, 639–702 (2019).
188. Wang, H. *et al*. RAVEN 2.0: A versatile toolbox for metabolic network reconstruction and a case study on Streptomyces coelicolor. *PLOS Comput. Biol*. **14**, e1006541 (2018).
189. Lieven, C. *et al*. MEMOTE for standardized genome-scale metabolic model testing. *Nature Biotechnology* (2020). doi:10.1038/s41587-020-0446-y
190. Courtot, M. *et al*. Controlled vocabularies and semantics in systems biology. *Mol. Syst. Biol*. **7**, 543 (2011).
191. Li, F. *et al*. Deep learning-based kcat prediction enables improved enzyme-constrained model reconstruction. *Nat. Catal*. **5**, 662–672 (2022).
192. Kanehisa, M., Furumichi, M., Sato, Y., Ishiguro-Watanabe, M. & Tanabe, M. KEGG: integrating viruses and cellular organisms. *Nucleic Acids Res*. (2020). doi:10.1093/nar/gkaa970
193. Monk, J. M. *et al*. iML1515, a knowledgebase that computes Escherichia coli traits. *Nature Biotechnology* **35**, 904–908 (2017).
194. Yang, L. *et al*. Principles of proteome allocation are revealed using proteomic data and genome-scale models. *Sci. Rep*. (2016). doi:10.1038/srep36734
195. Vikromvarasiri, N., Noda, S., Shirai, T. & Kondo, A. Investigation of two metabolic engineering approaches for (R, R)-2, 3-butanediol production from glycerol in Bacillus subtilis. *J. Biol. Eng*. **17**, 3 (2023).
196. Ng, C. Y., Jung, M. Y., Lee, J. & Oh, M. K. Production of 2,3-butanediol

in Saccharomyces cerevisiae by in silico aided metabolic engineering. *Microb. Cell Fact.* **11**, (2012).
197. Suástegui, M. *et al.* Multilevel engineering of the upstream module of aromatic amino acid biosynthesis in Saccharomyces cerevisiae for high production of polymer and drug precursors. *Metab. Eng.* **42**, (2017).
198. Ghiffary, M. R., Prabowo, C. P. S., Adidjaja, J. J., Lee, S. Y. & Kim, H. U. Systems metabolic engineering of Corynebacterium glutamicum for the efficient production of β-alanine. *Metab. Eng.* **74**, 121–129 (2022).
199. Xu, P., Ranganathan, S., Fowler, Z. L., Maranas, C. D. & Koffas, M. A. G. Genome-scale metabolic network modeling results in minimal interventions that cooperatively force carbon flux towards malonyl-CoA. *Metab. Eng.* **13**, 578–587 (2011).
200. Ranganathan, S. & Maranas, C. D. Microbial 1-butanol production: Identification of non-native production routes and in silico engineering interventions. *Biotechnol. J.* **5**, 716–725 (2010).
201. Cheng, F., Yu, H. & Stephanopoulos, G. Engineering Corynebacterium glutamicum for high-titer biosynthesis of hyaluronic acid. *Metab. Eng.* **55**, 276–289 (2019).
202. López, J. *et al.* Production of β-ionone by combined expression of carotenogenic and plant CCD1 genes in Saccharomyces cerevisiae. *Microb. Cell Fact.* **14**, (2015).
203. Chowdhury, A., Zomorrodi, A. R. & Maranas, C. D. k-OptForce: Integrating Kinetics with Flux Balance Analysis for Strain Design. *PLoS Comput. Biol.* (2014). doi:10.1371/journal.pcbi.1003487
204. Patil, K. R., Rocha, I., Förster, J. & Nielsen, J. Evolutionary programming as a platform for in silico metabolic engineering. *BMC Bioinformatics* **6**, 1–12 (2005).
205. Choi, H. S., Lee, S. Y., Kim, T. Y. & Woo, H. M. In silico identification of gene amplification targets for improvement of lycopene production. *Appl. Environ. Microbiol.* **76**, 3097–3105 (2010).
206. Moradi, S., Jahanian-Najafabadi, A. & Roudkenar, M. H. Artificial blood substitutes: first steps on the long route to clinical utility. *Clin. Med. Insights Blood Disord.* **9**, CMBD-S38461 (2016).
207. Fraser, R. Z., Shitut, M., Agrawal, P., Mendes, O. & Klapholz, S. Safety evaluation of soy leghemoglobin protein preparation derived from Pichia pastoris, intended for use as a flavor catalyst in plant-based meat. *Int. J. Toxicol.* **37**, 241–262 (2018).
208. Mardinoglu, A. *et al.* Genome-scale metabolic modelling of hepatocytes reveals serine deficiency in patients with non-alcoholic fatty liver disease. *Nat. Commun.* **5**, 1–11 (2014).
209. Hassing, E. J., de Groot, P. A., Marquenie, V. R., Pronk, J. T. & Daran, J. M. G. Connecting central carbon and aromatic amino acid metabolisms to improve de novo 2-phenylethanol production in Saccharomyces cerevisiae. *Metab. Eng.* **56**, (2019).
210. Larroude, M., Nicaud, J. M. & Rossignol, T. Yarrowia lipolytica chassis strains engineered to produce aromatic amino acids via the shikimate

pathway. *Microb. Biotechnol.* (2020). doi:10.1111/1751-7915.13745
211. Rajkumar, A. S. & Morrissey, J. P. Rational engineering of Kluyveromyces marxianus to create a chassis for the production of aromatic products. *Microb. Cell Fact.* **19**, (2020).
212. Qin, J. *et al.* Engineering yeast metabolism for the discovery and production of polyamines and polyamine analogues. *Nat. Catal.* (2021). doi:10.1038/s41929-021-00631-z
213. der Maaten, L. & Hinton, G. Visualizing data using t-SNE. *J. Mach. Learn. Res.* **9**, (2008).
214. O'Brien, E. J., Lerman, J. A., Chang, R. L., Hyduke, D. R. & Palsson, B. Genome-scale models of metabolism and gene expression extend and refine growth phenotype prediction. *Mol. Syst. Biol.* **9**, (2013).
215. Goelzer, A. *et al.* Quantitative prediction of genome-wide resource allocation in bacteria. *Metab. Eng.* **32**, (2015).
216. Bulović, A. *et al.* Automated generation of bacterial resource allocation models. *Metab. Eng.* (2019). doi:10.1016/j.ymben.2019.06.001
217. Christensen, T. S., Oliveira, A. P. & Nielsen, J. Reconstruction and logical modeling of glucose repression signaling pathways in Saccharomyces cerevisiae. *BMC Syst. Biol.* **3**, 7 (2009).
218. Romers, J., Thieme, S., Münzner, U. & Krantz, M. A scalable method for parameter-free simulation and validation of mechanistic cellular signal transduction network models. *npj Syst. Biol. Appl.* **6**, 2 (2020).
219. Yu, H. & Blair, R. H. Integration of probabilistic regulatory networks into constraint-based models of metabolism with applications to Alzheimer's disease. *BMC Bioinformatics* **20**, 386 (2019).
220. Postma, E., Verduyn, C., Scheffers, W. A. & P, V. D. J. Enzymic analysis of the crabtree effect in glucose-limited chemostat cultures of Saccharomyces cerevisiae. *Appl. Environ. Microbiol.* **55**, 468–477 (1989).
221. Park, J. O. *et al.* Near-equilibrium glycolysis supports metabolic homeostasis and energy yield. *Nat. Chem. Biol.* (2019). doi:10.1038/s41589-019-0364-9
222. Noor, E. *et al.* Pathway Thermodynamics Highlights Kinetic Obstacles in Central Metabolism. *PLOS Comput. Biol.* **10**, e1003483 (2014).

Paper I:

Stress-induced expression is enriched for evolutionarily young genes in diverse budding yeasts.

Doughty TW, Domenzain I, Millan-Oropeza, A., Montini, N., De Groot, P. A., Pereira, R., Nielsen, J., Henry, C., Daran, J. M., Siewers, V., Morrissey, J. P.

Nature Communications 2020

nature COMMUNICATIONS

ARTICLE

https://doi.org/10.1038/s41467-020-16073-3 OPEN

Stress-induced expression is enriched for evolutionarily young genes in diverse budding yeasts

Tyler W. Doughty[1,2], Iván Domenzain[1,2], Aaron Millan-Oropeza[3], Noemi Montini[4], Philip A. de Groot[5], Rui Pereira[1,2], Jens Nielsen[1,2], Céline Henry[3], Jean-Marc G. Daran[5], Verena Siewers[1,2]✉ & John P. Morrissey[4]✉

The Saccharomycotina subphylum (budding yeasts) spans 400 million years of evolution and includes species that thrive in diverse environments. To study niche-adaptation, we identify changes in gene expression in three divergent yeasts grown in the presence of various stressors. Duplicated and non-conserved genes are significantly more likely to respond to stress than genes that are conserved as single-copy orthologs. Next, we develop a sorting method that considers evolutionary origin and duplication timing to assign an evolutionary age to each gene. Subsequent analysis reveals that genes that emerged in recent evolutionary time are enriched amongst stress-responsive genes for each species. This gene expression pattern suggests that budding yeasts share a stress adaptation mechanism, whereby selective pressure leads to functionalization of young genes to improve growth in adverse conditions. Further characterization of young genes from species that thrive in harsh environments can inform the design of more robust strains for biotechnology.

[1] Department of Biology and Biological Engineering, Chalmers University of Technology, SE-41296 Gothenburg, Sweden. [2] Novo Nordisk Foundation Center for Biosustainability, Chalmers University of Technology, SE-41296 Gothenburg, Sweden. [3] Plateforme d'Analyse Protéomique Paris Sud-Ouest (PAPPSO), INRAE, MICALIS Institute, Université Paris-Saclay, 78350 Jouy-en-Josas, France. [4] School of Microbiology, Environmental Research Institute and APC Microbiome Ireland, University College Cork, Cork T12YN60, Ireland. [5] Department of Biotechnology, Delft University of Technology, Van der Maasweg 9, 2629 HZ Delft, The Netherlands. ✉email: siewers@chalmers.se; j.morrissey@ucc.ie

Yeasts in the Saccharomycotina subphylum, (budding yeasts), have proven to be useful platforms for the production of ethanol, flavors, nutritional supplements, biopharmaceuticals, as well as other valuable chemicals[1–3]. At present, industrial production using budding yeasts is dominated by the extensively characterized species *Saccharomyces cerevisiae*. *S. cerevisiae* exhibits common budding yeast phenotypes (e.g., efficient growth on some simple sugars) as well as a less common adaptation amongst budding yeasts, high ethanol tolerance[4]. Together, these traits enable cost-effective production of 100 billion liters of ethanol annually using *S. cerevisiae* as a platform[1]. Other budding yeasts have adaptations that make them well-suited for production of specific biomolecules, something that is possible due to the improved strain engineering capacity following the emergence of CRISPR/Cas9[5,6]. Examples are *Yarrowia lipolytica*, which evolved to tolerate hydrophobic environments and can produce high-yields of fatty acids[7,8], and *Kluyveromyces marxianus*, whose thermotolerance is a beneficial feature for industrial processes[6,9]. Despite progress in sequencing genomes and phenotypic characterization of these and many other yeast species, the genes that underpin adaptation to cope with harsh conditions remain enigmatic.

For the species above, adaptations to natural environments enable robustness in industrial biotechnology processes. Understanding the genes that influence these and other exceptional stress tolerances would enable the engineering of more robust industrial strains, thereby reducing process costs and increasing yields[10,11]. Although studies that sought to characterize stress tolerances in *S. cerevisiae* have elucidated mechanisms that influence robustness[10,12,13], engineering more robust *S. cerevisiae* strains without physiological trade-offs remains challenging[9]. One complication is that stress exposure often results in hundreds of significant transcriptional changes[13,14], most of which do not correlate with single gene deletion changes in robustness[11]. These results suggest that multiple genes from different gene families may contribute additively to robustness and/or that stress genes may exist as duplicates, as is the case for antifreeze protein genes in artic yeasts[15]. Thus, researchers have employed systems biology to characterize the transcriptome and/or proteome-wide stress-induced changes[13,14,16–18]. These approaches have identified biological processes that exhibit altered expression in response to stress exposure, which builds upon and relates to previous research into gene functions (e.g., GO term enrichment analysis). These associations are possible due to extensive annotations of *S. cerevisiae* genes that result from decades of experimental analyses[19]. For most other yeast species, the majority of gene functional information is acquired second hand via homology search tools. This paradigm results in a large portion of genes of unknown function, which is especially large for species that are phylogenetically distant from extensively characterized species like *S. cerevisiae*[20]. These uncharacterized genes are difficult to integrate into omics analyses like GO term enrichment, as they do not have a known function or localization. Because of this, gene functional analysis of poorly characterized species is restricted to conserved genes, which may not be the only genes that influence stress-tolerance phenotypes. Currently, hundreds of whole genome sequences are available from diverse budding yeasts[21], including several species that are known to exhibit extreme stress tolerances[22], but many of the causative genes that enable yeast stress tolerances remain elusive.

Here, we analyze stress conditions to assess gene expression changes after stress adaptation in three diverse budding yeast species, one of which is well characterized (*S. cerevisiae*), and two that are less-well-characterized (*K. marxianus* and *Y. lipolytica*). The goal of this analysis is to identify common systems-level trends that are shared between each species stress responses. This analysis discovers that each organism displays a consistent response at the level of gene expression that is characterized by the enrichment of stress responsive genes amongst certain categories: namely, genes of unknown function and recently (in evolutionary time) duplicated and taxonomically restricted genes (young genes). The findings of this work suggest an evolutionary mechanism that is biased for stress tolerance functionalization and stress-induced expression of young genes. We propose that the gene sorting method we developed provides a path forward for more rapid identification of stress response genes in environmentally robust yeast, thereby accelerating understanding of niche adaption in budding yeasts.

Results

Conserved category enrichment of stress responsive genes. In this work, *S. cerevisiae*, *K. marxianus*, and *Y. lipolytica* were exposed to stress conditions that are present in natural environments, such as those caused by environmental temperature variation and growth on sugar-rich or acidic substrates[22]. These stress responses are also industrially-relevant, as they are caused by feedstocks (high osmotic pressure and low pH) or process conditions (elevated temperatures) during industrial fermentations[11]. Characterizing stress responses in these species is valuable due to their phylogenetic diversity, which spans much of the Saccharomycotina subphylum[21]. To minimize noise caused by variable growth rate[23], experiments were carried out in steady-state chemostats at a fixed growth rate under standard and stress conditions. This experimental setup allows strains to adjust to the conditions imposed by sub-lethal stress before sampling and analysis. Transcriptomic changes that occurred in response these stress conditions were identified via differential expression analysis (Fig. 1a).

To understand the function of stress responsive genes, biological process annotations were acquired from Ensembl (*S. cerevisiae*) or identified using BLAST2GO[20] for (*K. marxianus* and *Y. lipolytica*). BLAST2GO annotated gene functions to otherwise unknown genes based on homology to an experimentally characterized gene. This process failed to annotate 20% and 38% of the mRNAs measured by RNAseq in this study for *K. marxianus* and *Y. lipolytica*, respectively (Supplementary Fig. 1A). The lower frequency of gene annotation for *Y. lipolytica* was expected, since this species is not closely related to extensively characterized yeasts[21]. Comparison of gene annotations and differential gene expression showed a higher percentage of genes of unknown function that were stress responsive than would be expected. For example, 38% of all protein-coding genes measured in this study for *Y. lipolytica* lacked a functional annotation, while 50% of stress responsive genes were genes of unknown function (Supplementary Fig. 1B).

This high proportion of stress-responsive genes of unknown function suggested that the most broadly conserved genes, which often have functional annotations, might be under-represented amongst the stress responses. To assess this, orthologous proteins shared between the three yeast species were inferred using OrthoFinder, which enables proteome-wide matching based on amino-acid sequence and chain length similarity in order to predict proteins that descend from a common ancestor[24]. To assess the fidelity of ortholog predictions, protein complexes and enzymatic processes that were previously characterized as conserved amongst budding yeasts as single-copy genes were searched for amongst orthology inference results[25]. This analysis found that orthology inference identified the majority of the expected complex members and enzymes as orthologs (Supplementary Fig. 2B), which supports the high fidelity of OrthoFinder

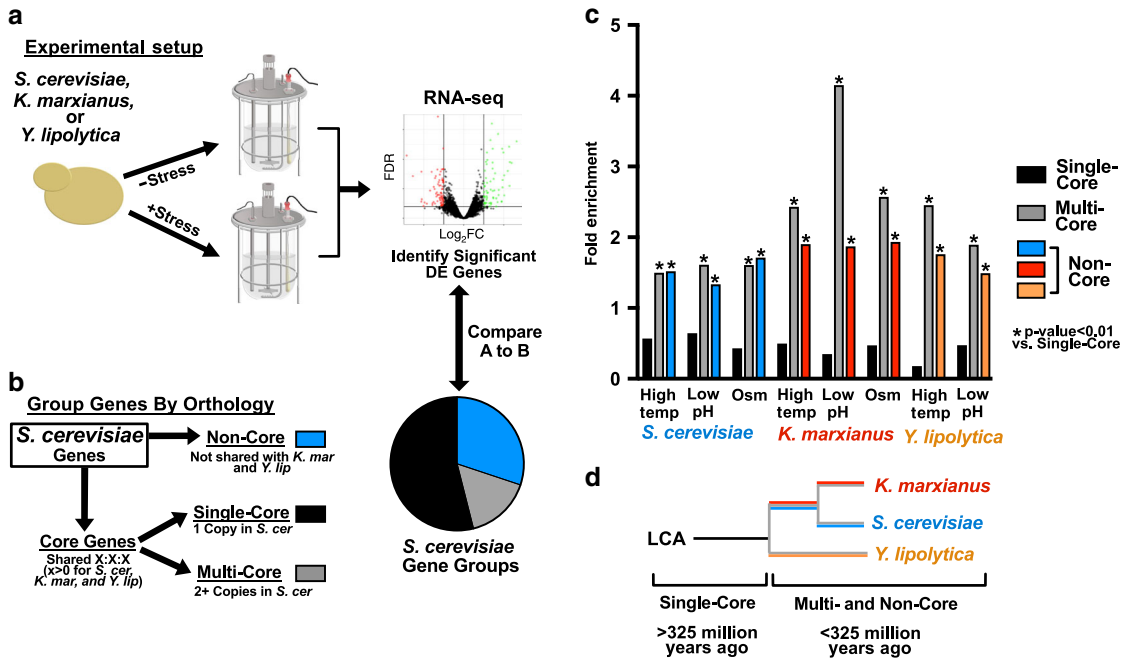


Fig. 1 Stress adaptation responsive genes are enriched for duplicated and non-conserved genes. a *S. cerevisiae*, *K. marxianus*, and *Y. lipolytica* were cultivated in chemostats in standard conditions or in the presence of stress (elevated temperature, low pH, or KCl). RNAseq was performed followed by differential expression analysis. **b** The protein-coding genes of each organism were compared to infer orthology using OrthoFinder. The resulting gene groups for *S. cerevisiae* are shown, with single-copy orthologous genes (Single-Core [black]), multi-copy orthologous genes (Multi-Core [gray]), and genes that were not shared (Non-Core [blue]). **c** The number of differentially expressed ($log_2FC > 1$, FDR < 0.01) mRNAs were divided by the total number of detected mRNAs inside of each ortholog group. Values were normalized to the overall DE gene # divided by the total genes measured, *p*-values were calculated using a two-sided Fisher's exact test. **d** A simplified phylogenetic tree. Single Core orthologs are predicted to originate from a Last Common Ancestor >325 million years ago. Multi- and Non-Core Genes are predicted to have duplicated or arisen de novo <325 million years ago.

predictions that was observed previously[24]. The results of the orthology inference analysis were used to divide each protein into one of three classes, single-core orthologous, multi-core orthologous, and non-orthologous. These proteins were matched to their corresponding genes for comparison to RNAseq differential expression. Gene sorting examples are shown in Supplementary Fig. 2A and the complete lists of genes for *S. cerevisiae, K. marxianus* and *Y. lipolytica* are in Supplementary Data 1, Supplementary Data 2 and Supplementary Data 3, respectively.

The results of orthology inference for *S. cerevisiae* are shown in Fig. 1b as an example. Each measured protein-coding gene from *S. cerevisiae* was identified as either (1) present as a single-copy gene with an ortholog in *K. marxianus* and *Y. lipolytica* (black Single-Core), (2) present as a duplicated gene with an ortholog in *K. marxianus* and *Y. lipolytica* (gray Multi-Core), or (3) lacking an ortholog in *K. marxianus* or *Y. lipolytica* (color Non-Core). The resulting groups were compared to the observed differentially expressed (DE) genes, which showed that multi-core and non-core genes were significantly enriched amongst DE genes in each stress condition tested (Fig. 1c). The same gene sorting regime shows that *K. marxianus* and *Y. lipolytica* exhibited similar DE gene enrichment for the multi-core and non-core gene groups (Fig. 1c and Supplementary Fig. 3A). Similar results were found amongst proteomics measurements for some stress conditions (Supplementary Methods 2–5), but this analysis was hindered by low detection of non-core proteins (Supplementary Fig. 3C).

The phenomenon depicted in Fig. 1C shows that single-core genes, which are predicted to have descended from a last common ancestor between the three yeast species (approximately 325 million years ago[21]), were under-represented amongst stress responsive genes for each stress and each organism. In contrast, genes that have duplicated or emerged in more recent evolutionary time were enriched amongst stress responsive genes. These observations suggest that evolutionary events may predict differential expression amongst these diverse yeast species (Fig. 1d).

***S. cerevisiae* stress response is enriched for young genes.** The results in Fig. 1 suggested a relationship between the genes that exhibit differential expression in response to stress and evolutionary events, like de novo gene emergence and gene duplication. Further characterization of this relationship could aid in understanding stress gene evolution and could help to predict genes that enable stress tolerance. Thus, we sought to test this relationship more stringently by dividing the protein-coding genes of *S. cerevisiae* into more precise groups that collectively represent a broad swath of eukaryotic evolution. The resulting groups are referred to as gene age groups, which were determined by ortholog presence at shared copy number in common ancestors that date from over 400 million years ago to 20 million years ago[21]. A similar approach, phylostratigraphy, divides genes into groups based on homology and has been used to infer gene origination events to identify periods in evolution that correlate with adaptive events[26]. However, the results in Fig. 1c indicated that an analysis procedure that considers both gene origin timing (like phylostratigraphy) and gene duplication timing could provide insights into stress responsive gene expression.

Gene grouping based on gene age was assessed using OrthoFinder[24] and is described in detail in Supplementary Method 1. Briefly, all *S. cerevisiae* genes were divided into three initial subsets; (1) fixed duplicates from the whole-genome duplication (WGD)[27], (2) genes that are present as single-copy genes, and (3) duplicate genes that arose outside of the whole-genome duplication (non-WGD) (Supplementary Fig. 4A).

Ortholog inference was used to sort each of the 4351 single-copy genes into a single bin based on the most distant ancestor with an orthologous gene using the hierarchal approach shown in Supplementary Fig. 4C. The multi-copy non-WGD gene groups were sorted by the presence of orthologous genes with the same copy number in a bottom-up approach in order to trace the relative timing of gene duplication events (Supplementary Fig. 4D). Finally, genes that were duplicated during the whole-genome duplication were grouped together. This sorting method matched each protein coding gene from *S. cerevisiae* to a single group that reflects the timing of the emergence (single-copy genes) or timing of duplication (multi-copy genes) of each gene, which we refer to as gene age. The inherent limitation with this approach is the availability of accurately annotated genome sequences across the phylogenetic tree. In the future, more phylogenetic information and additional gene matching algorithms will improve the fidelity of gene age prediction and may lead to some refining of the gene age categorization. Gene sorting examples are shown in Supplementary Fig. 2A and the complete list of genes can be found in Supplementary Data 4.

The gene groupings in Fig. 2b were compared to the stress RNAseq data to determine the percentage of significantly differentially expressed genes in each age group. This analysis found a stepwise increase in the relative amount of differentially expressed genes in progressively younger gene groups in *S. cerevisiae*. Genes that were found to be conserved to filamentous fungi (ancient genes from group I) were 4.2 to 6.6-fold less likely to be differentially expressed after stress adaptation compared to *S. cerevisiae*-specific genes (group V) (Fig. 2c). Similar trends were observed when considering only upregulated or downregulated genes, however, upregulated genes showed a more pronounced bias towards young genes with 6.6 to 16.8-fold enrichment between group I and group V genes (Supplementary Fig. 5). Analysis of the expression pattern of young genes (those in groups IV and V) showed that few genes exhibited significantly changed expression in response to all stresses (Fig. 2d, e).

The findings in Fig. 2 were further tested by analyzing additional stress adaptation experiments for *S. cerevisiae* exposed to ethanol in a previous study[28] or anaerobic stress (this study) (Supplementary Fig. 6). In both cases, young genes were enriched, and ancient genes were depleted amongst differentially expressed genes in response to stress adaptation. A similar enrichment for young genes was observed amongst varying amounts of ethanol stress, despite a difference in the number of total significant gene expression changes (Supplementary Fig. 6D). Together, these observations suggest that the sorting algorithm presented in Supplementary Fig. 4 is able to consistently identify a relationship between gene age and stress gene expression for several types and levels of stress in *S. cerevisiae*.

Shared gene enrichment pattern across the Saccharomycotina. The findings in Fig. 2 showed an inverse correlation between gene age and stress differential expression in *S. cerevisiae*. If these findings were shared amongst other yeast species, they might imply an underlying evolutionary mechanism that can predict the genes that are more likely to be involved in stress adaptation. To test for a relationship between differential expression and gene age, we stratified the protein-coding genes of *K. marxianus* and *Y. lipolytica* using the same sorting concept described above for *S. cerevisiae* (Supplementary Fig. 4). The only modification to these sorting approaches was the elimination of the whole-genome

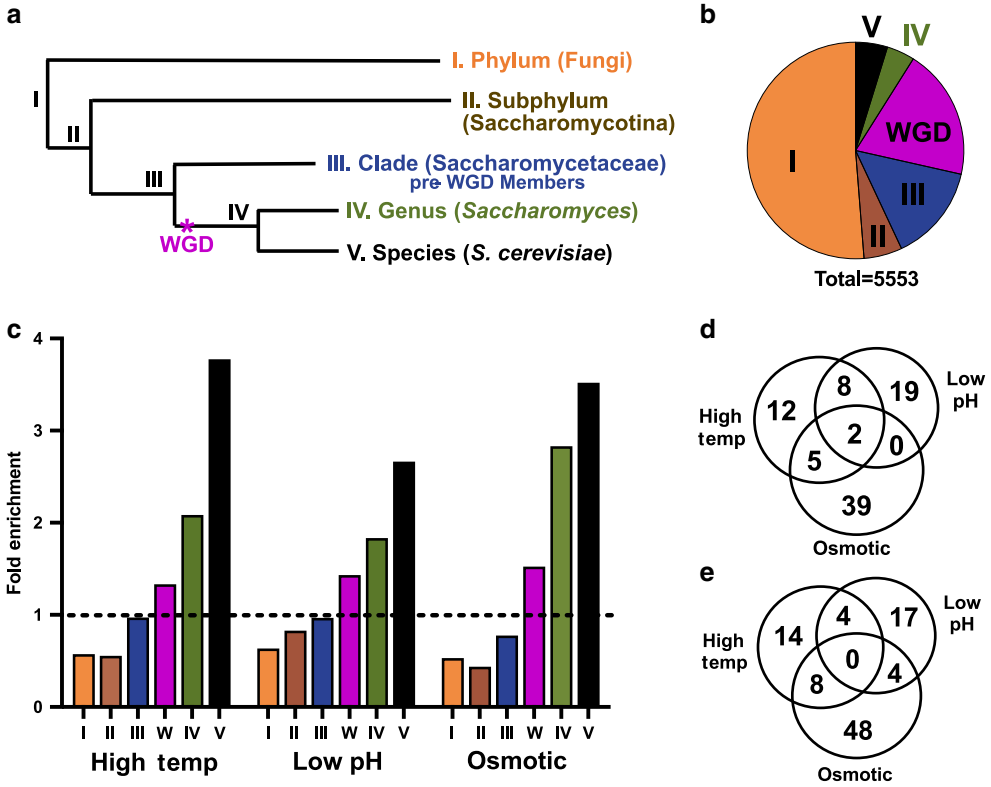


Fig. 2 Stress adaptation responsive genes in *S. cerevisiae* are enriched for young genes. a A simplified phylogenetic tree for *S. cerevisiae* showing speciation events and the Whole Genome Duplication (magenta*). **b** The transcripts detected via RNAseq from this study were grouped based on ortholog presence in the groups shown (described in detail in Supplementary Fig. 4). **c** Differentially expressed genes for *S. cerevisiae* were parsed by their grouping shown in **b**, then normalized to the group size and the proportion of total Differentially Expressed (DE) genes per condition (dashed line). Transcripts in groups IV and V were assessed for shared upregulated genes (D) or downregulated genes (E).

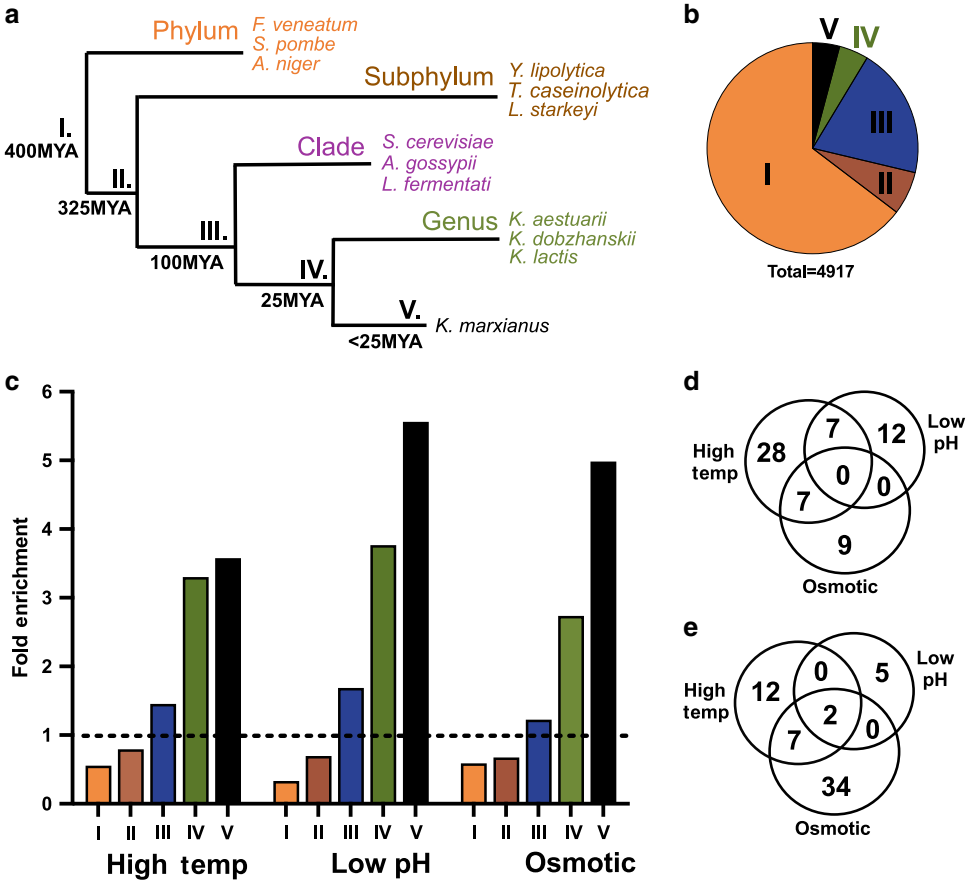


Fig. 3 Stress adaptation responsive genes in *K. marxianus* are enriched for young genes. a A simplified phylogenetic tree for *K. marxianus* showing speciation events and organisms used in orthology queries. **b** The transcripts detected via RNAseq from this study were grouped based on ortholog presence in the groups shown (described in detail in Supplementary Fig. 4). **c** Differentially expressed genes for *K. marxianus* were parsed by their grouping shown in **a** and **b**, then normalized to the group size and the total measured DE % (dashed line). Transcripts in groups IV and V were assessed for shared upregulated genes (**d**) or downregulated genes (**e**).

duplication group, as neither of these species has undergone a recent whole-genome duplication[29,30].

Analysis of *K. marxianus* and *Y. lipolytica* gene groups in relation to each stress condition showed similar patterns to *S. cerevisiae*, with ancient genes exhibiting under-representation for significant differential expression compared to young gene groups (Fig. 3 and Supplementary Fig. 7). Also, as with *S.* cerevisiae, there were few young differentially expressed genes that responded to all stresses, suggesting that these expression changes were often condition specific (Fig. 3d, e). These biases towards young genes might explain the low observed overlap between significant expression changes amongst 1:1:1 orthologs shared between the three budding yeasts when exposed to the same type of stress (Supplementary Fig. 8). Together, these findings showed that in all three yeasts studied, young genes were enriched for long-term stress-responsiveness, or adaptation, compared to ancient genes. Further, since the species chosen for this analysis span much of the diversity of the budding yeast subphylum[21], these results may be indicative of a shared stress adaptation mechanism, rather than a shared response of specific genes, amongst budding yeasts.

Features of young genes are consistent with adaptive roles. To understand the functions associated with the gene groupings produced in this study, we assessed biological processes associated with the ancient and young gene sets in *S. cerevisiae*, where ample functional information is available. This analysis showed ancient genes associated with fundamental biological processes including primary metabolism, tRNA aminoacylation, and DNA strand elongation, and 94% of these genes were annotated with at least one biological process GO term. Conversely, young genes (groups IV and V) were associated with more specialized functions like maltose transport, vitamin biosynthesis, and aldehyde metabolism, with many young genes lacking any biological process annotations in *S. cerevisiae* (40%). *K. marxianus* and *Y. lipolytica* also exhibited high percentages of young genes that were not associated with a biological process (41% and 69%, respectively) (Supplementary Fig. 9B). The fundamental nature of ancient gene functional associations was reflected by their high likelihood of being essential or required for optimal growth compared to young genes. Conversely, the more specialized functions of young genes were reflected by the 16-fold decrease in likelihood of growth impairment upon deletion compared to ancient genes (Fig. 4c)[31]. Analysis of cellular component enrichment showed that young proteins (groups IV and V) were significantly enriched for localization to the plasma membrane, cell wall, and vacuole, which was distinct from ancient proteins (group I) enrichment for nuclear, cytoplasmic, and mitochondrial localization (Supplementary Fig. 9B).

Further characterization of young protein-coding genes found that they exhibited lower median gene expression and their corresponding proteins were less frequently detected via mass spectrometry in non-stress samples compared to ancient genes (Figs. 4a, b). Previous works have shown that low expression and non-essentiality correlate with increased adaptation rates[32,33],

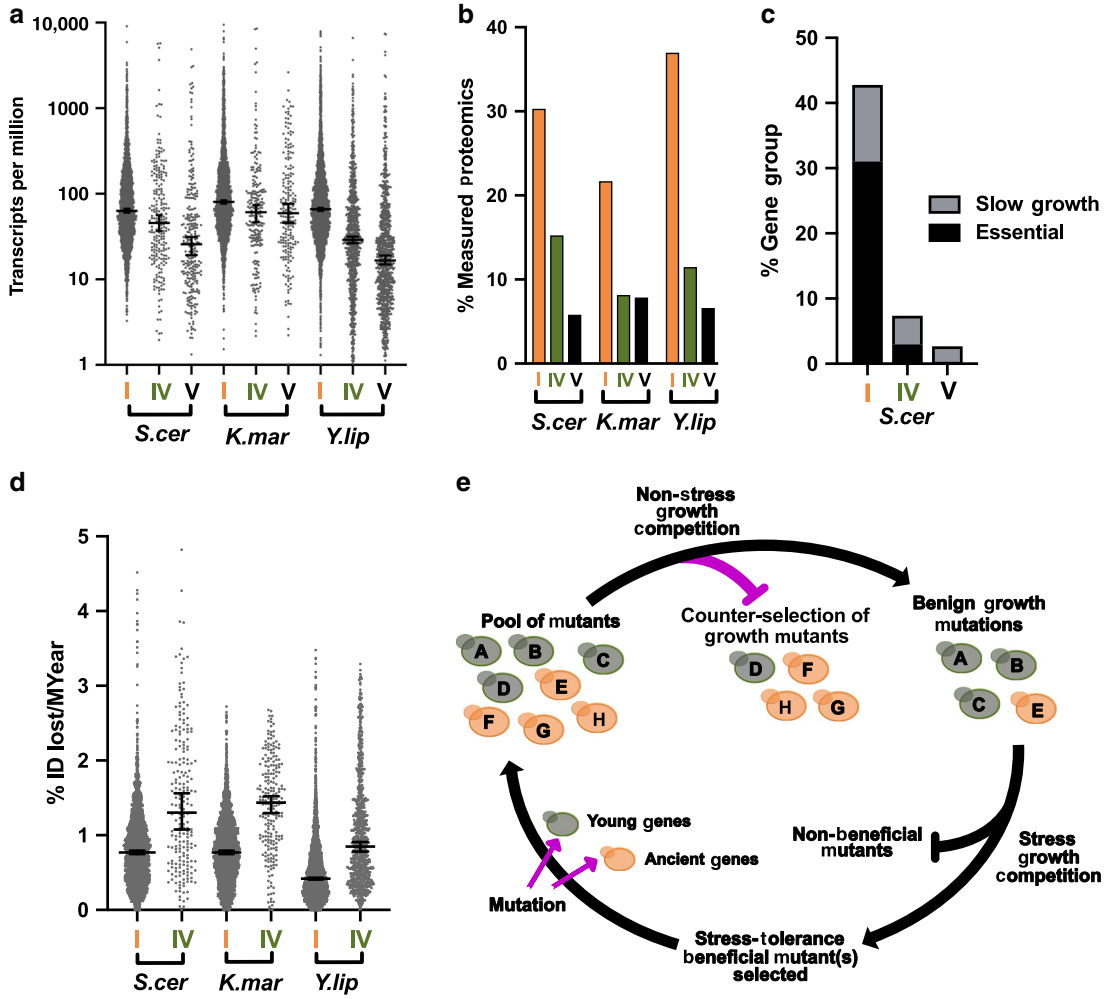


Fig. 4 Less expressed and often non-essential young genes adapt more rapidly than ancient genes. a Standard growth condition RNAseq reads were normalized to the read depth and gene length to generate Transcripts per Million (TPM). Error bars at the 95% confidence interval of the median. **b** The percentage of mRNAs measured compared to proteins measured via mass spectrometry by quantifying eXtracted Ion Chromatograms. **c** The percentage of essential genes (black) and non-essential genes associated with slow growth (gray) is shown for *S. cerevisiae* ancient genes (I) and young genes (IV and V). Essential and slow growth ORFs were obtained from Giaever 2002[20]. **d** The percentage of amino acid identity changes for each protein in comparison to its closest homolog from a member of the same genus. Results were adjusted to the percent amino acid change per million years (% Intentity (ID) lost/MYear) using the estimated divergence time between pairs of organisms[13]. The median and 95% confidence interval is shown. Queries were performed between *S. cerevisiae/S. eubayanus*, *K. marxianus/K. lactis*, or *Y. lipolytica/Y. bubula*. E. A model for evolution to intermittent stress where random mutations occur amongst all genes (magenta arrows) followed by non-stress selection for benign mutants (magenta blocked arrow). Mutants that do not influence growth are selected upon stress exposure for fitness benefits. Source data underlying Fig. 4a, c, and d are provided as a source data file.

suggesting that young genes could adapt more rapidly compared to ancient genes. To test this, amino acid sequence identity was compared between homologous proteins from members of the same genus using BLAST+. Analysis of each protein sequence from groups I and IV allowed sequence identity changes to be compared over the same span of evolutionary time to assess adaptation rates. This analysis was adjusted to reflect the estimated evolutionary time elapsed[21] between each pair of species and showed that the average frequency of amino acid identity changes was higher for young protein groups compared to ancient protein groups (Fig. 4d).

Discussion

Budding yeasts are attractive for industrial production of biomolecules, since they grow rapidly, utilize inexpensive substrates, and are readily engineered to produce heterologous gene products[1–3]. However, stresses that result from feedstock composition, toxic products, and fluctuating reaction temperatures can lower the cost-effectiveness of industrial processes by diminishing productivity and yields[11]. Previous works have phenotypically characterized yeasts exhibiting stress tolerant phenotypes[22], and whole genome sequencing data are available, but the genes that have evolved in these yeasts to enable survival and growth under unfavorable, stress-inducing conditions remain unclear. We now identify an association between stress-induced gene expression and gene age. We show that younger genes, namely, those that are restricted to a genus or species, or have duplicated in recent evolutionary time, are more likely to respond to different types of long-term stress, such as those that were imposed in continuous (chemostat) cultivation in this report. These stress-responsive genes can also be considered adaptation or niche-specialization genes as they have evolved to enable the yeasts carrying them tolerate ongoing harsh conditions.

The findings that adaptation rates and stress gene expression are biased toward young genes for three distantly related yeast species suggests an underlying evolutionary mechanism. The

model in Fig. 4e suggests that during non-stress periods, ancient and young gene mutations may occur at similar rates, however, ancient genes may be subject to more stringent counter-selection (magenta blocked arrow) due to their higher expression and influence on growth (Fig. 4a, c). Conversely, non-synonymous mutations amongst young genes might accumulate more rapidly because these genes are rarely growth-related (Fig. 4c, d). The resulting increase in sequence space that is sampled by young genes would increase the probability of young mutants to enter stress-growth competition, thus increasing the chances of selecting young gene adaptations to benefit stress tolerance. We suggest that these events occur in a cyclical manner, enabling stress-tolerance functionalization of young genes without diminishing growth potential. This model could also apply to promoter sequences, which would enable specialized genes to adapt dynamic expression patterns in order to save resources during non-stress growth. This mechanism would explain the higher propensity of young genes to change expression in response to stress. The model might also provide an insight as to why improved stress tolerance in some laboratory-evolved strains comes at a cost to growth under standard growth conditions[34,35]. In this case, the relatively short, non-cyclical stresses applied during adaptive laboratory evolution does not allow for the counterselection of growth mutations.

In this work we found that young genes represented 4%, 5%, and 14% of protein-coding genes in *K. marxianus, S. cerevisiae*, and *Y. lipolytica*, respectively, which is in the same range as the 7-19% of genes in *C. elegans, D. melanogaster*, and *H. sapiens* that lack recognizable homologs in other organisms[26,36]. Previous works have linked some young genes to species and genus-specific adaptations, including movement on the surface of fast water in *Rhagovelia* water striders[37], HIV-1 resistance in owl monkeys[38,39], and the concurrent evolution of antifreeze proteins in several species[40–42]. Antifreeze protein genes are well-studied examples of young genes that arose via de novo gene origin events between 13 and 18 million years ago in codfishes and are present at variable copy number in some species[43]. Concurrently, the psychrophilic yeast *G. antarctica*, has evolved to encode nine antifreeze protein genes whose expression levels are induced by exposure to cold[15,44]. These attributes of antifreeze protein genes are similar to the young genes in this study, which were stress responsive, emerged in recent evolutionary time, and often exist at variable copy number. It seems plausible that the young, stress responsive genes described for *K. marxianus* could influence the capacity of this species to grow at higher temperatures (45 °C)[9] than other members of the *Kluyveromyces* genus, like *K. lactis* (37 °C)[45]. Furthermore, the acquisition of this thermotolerant phenotype in a short span of evolutionary time would be consistent with the involvement of rapidly adapting young genes.

This study and previous stress tolerance investigations have identified dozens to hundreds of significant gene expression changes after stress exposure in budding yeasts[13,16–18,28]. Despite analysis of such stress-responsive genes in multiple species, rational engineering to further enhance robustness of industrial yeast strains remains difficult. The findings of this work suggest that considering the collective role of evolutionarily young stress-responsive genes from stress tolerant species is a pragmatic path forward towards achieving this goal. This suggestion is based on two points; first, single gene perturbations often fail to reproduce stress-response phenotypes[13]; and second, many mutations that improve stress tolerance cause trade-off phenotypes[10,34,35]. Establishing more robust industrial production strains may require modification of multiple genes and/or expression of several exogenous genes, while avoiding growth or physiological perturbations. To accomplish this, knowledge-driven approaches are needed to aid the identification of relevant genes that can be manipulated to confer the desired trait without negative consequences on growth. This goal is complicated by incomplete gene function information, especially for many stress tolerant yeast species. In this work, we present a gene sorting method that identifies a class of genes that are likely to be enriched in response to diverse stresses. By leveraging gene age information, it will be possible to focus rational experimental designs on unpredicted stress tolerance genes, which prior to this work fall into the category of genes of unknown function. Identifying these genes using this analysis methodology offers biotechnological potential as well as the tools to understand the process of species diversification and niche adaptation in yeast.

Methods

Strains and cultivation conditions. *Y. lipolytica* (W29), *K. marxianus* (CBS6556), and *S. cerevisiae* (CEN.PK113-7D) were grown in 30 mL synthetic media at 30 °C for 24 h in shake flasks, followed by inoculation of bioreactors and an initial batch growth phase. After the completion of the batch phase, chemostat cultivation was started with a dilution rate of 0.1/h and a working volume of 500 mL (*S. cerevisiae*) or 1 L (*K. marxianus* and *Y. lipolytica*). Stress conditions were achieved by altering either temperature, pH, or osmotic pressure (KCl) for the duration of the cultivation, specific conditions are listed in Supplementary Fig. 8. Standard growth temperature was adjusted to reflect organism specific tolerances. Cultivations for were performed in synthetic medium (SM)[46] containing 5 g L^{-1} $(NH_4)_2SO_4$, 3 g L^{-1} KH_2PO_4, 0.5 g L^{-1} $MgSO_4{\cdot}7H_2O$, 7.5 g L^{-1} glucose, trace elements and vitamins with 1 g L^{-1} pluronic PE6100 to reduce foaming. Sample collection was carried out after at least five volume changes (50 h) in steady state growth conditions. At least three biological replicate experiments were performed for each species and each condition in this work. Steady state growth was defined as less than 5% deviation in biomass dry weight.

Ortholog prediction with OrthoFinder. For Fig. 1, proteome-wide homology matching was executed using OrthoFinder[24]. Proteins were excluded from the core genome (non-core) if orthology search predicted zero orthologous proteins in any of the query species. Proteins were designated single-core if they were encoded by single-copy genes in the species (e.g., *S. cerevisiae HIS1*) or multi-core if they were duplicated in the species (e.g., *S. cerevisiae GAL1* and *GAL3*) (Supplementary Fig. 2). Protein groups were matched to their underlying genes for gene expression analyses. This grouping strategy was carried out to sort each species protein-coding genes into a single group. Results of these gene sorting analyses are shown in Supplementary Data 1, Supplementary Data 2 and Supplementary Data 3. For Figs. 2 and 3, and Supplementary Fig. 7, OrthoFinder was used to identify orthologs between each yeast and a set of eukaryotic organisms. This is shown in Supplementary Fig. 4 and is discussed in more detail in Supplementary Method 1. The results of these gene sorting analyses are shown in Supplementary Data 4, Supplementary Data 5 and Supplementary Data 6.

RNAseq preparation and mapping. RNA extractions were performed on samples that were mechanically lysed with 0.5 mm acid-washed beads using an MP-Biomedicals FastPrep-24 for three one-minute cycles. Further extraction was performed using an RNeasy Kit from Qiagen. Libraries were prepared using the TruSeq mRNA Stranded HT kit. Sequencing was carried out using an Illumina NextSeq 500 High Output Kit v2 (75 bases), with a minimum of 8 million paired-end reads per replicate. The Novo Nordisk Foundation Centre for Biosustainability (Technical University of Denmark), performed the RNA sequencing and library preparation. RNAseq read mapping was performed after analysis in FASTQC, which identified one sample from *K. marxianus* as having overrepresented sequences. This sample was excluded from the analysis herein. Analysis for TPM in Fig. 4a was performed using Hisat2 v2.1.0[47] and StringTie v1.3.3b[48]. RNAseq mapping for differential expression was mapped with STAR v2.7.0[49] and reads were assigned with featureCounts v1.6.0[50]. Differential expression results can be found in Supplementary Data 1, Supplementary Data 2 and Supplementary Data 3.

Differential expression analysis. Differential expression results were generated using limma v3.40.6[51] and edgeR v3.26.8[52] R packages and tidyverse v1.3.0[53] was employed for various data rearrangements. Filtering was used to remove lowly expressed genes/proteins, and each dataset was filtered to remove genes/proteins for which the relative standard deviation was greater than 1 (RSD > 1) across replicates for a given condition and organism. Differential expression was defined by a significance cutoff of absolute $\log_2$FC > 1 and False Discovery Rate < 0.01 for a stress condition compared to control. The data analysis pipeline is described in Supplementary Method 6.

Reporting Summary. Further information on research design is available in the Nature Research Reporting Summary linked to this article.

Data availability

Data supporting the findings of this work are available within the paper and its Supplementary Information files. A reporting summary for this Article is available as a Supplementary Information file. All mapped transcript data and protein detection data generated in this work can be found at https://github.com/SysBioChalmers/OrthOmics. RNAseq datasets of data generated in this study can be found using SRA accession PRJNA531619 [https://www.ncbi.nlm.nih.gov/bioproject/PRJNA531619/]. Additional RNAseq data analyzed in Supplementary Fig. 6 are available in the ArrayExpress database with the dataset ID E-MTAB-4044 [https://www.ebi.ac.uk/arrayexpress/experiments/E-MTAB-4044/]. Proteomics data is available via the PRIDE partner repository with the dataset ID PXD011426 [http://proteomecentral.proteomexchange.org/cgi/GetDataset?ID=PXD011426]. The source data underlying Figs. 4a, c, and d, as well as Supplementary Figs. 1A, 2B, 3B, 3C, 6B, 6D, and 9 are provided as a Source Data file.

Code availability

All custom tools and analysis scripts can be freely accessed at github repository [https://github.com/SysBioChalmers/OrthOmics].

Received: 28 May 2019; Accepted: 9 April 2020;
Published online: 01 May 2020

References

1. Mohd Azhar, S. H. et al. Yeasts in sustainable bioethanol production: a review. *Biochem. Biophys. Rep.* **10**, 52–61 (2017).
2. Nielsen, J. & Keasling, J. D. Engineering cellular metabolism. *Cell* **164**, 1185–1197 (2016).
3. Sanchez-Garcia, L. et al. Recombinant pharmaceuticals from microbial cells: a 2015 update. *Microb. Cell Fact.* **15**, 33 (2016).
4. Ma, Y. J., Lin, L. L., Chien, H. R. & Hsu, W. H. Efficient utilization of starch by a recombinant strain of *Saccharomyces cerevisiae* producing glucoamylase and isoamylase. *Biotechnol. Appl. Biochem.* **31**, 55–59 (2000).
5. Ledesma-Amaro, R. & Nicaud, J. M. *Yarrowia lipolytica* as a biotechnological chassis to produce usual and unusual fatty acids. *Prog. Lipid Res.* **61**, 40–50 (2016).
6. Varela, J. A., Gethins, L., Stanton, C., Ross, P. & Morrissey, J. P. Applications of *Kluyveromyces marxianus* in biotechnology. In *Yeast Diversity in Human Welfare* (eds. Satyanarayana, T. & Kunze, G.) 439–453 (Springer, Singapore, 2017). https://doi.org/10.1007/978-981-10-2621-8_17
7. Nicaud, J.-M. Yarrowia lipolytica. *Yeast* **29**, 409–418 (2012).
8. Gonçalves, F. A. G., Colen, G. & Takahashi, J. A. *Yarrowia lipolytica* and its multiple applications in the biotechnological industry. *Sci. World J.* **2014**, 476207 (2014).
9. Lane, M. M. & Morrissey, J. P. *Kluyveromyces marxianus*: a yeast emerging from its sister's shadow. *Fungal Biol. Rev.* **24**, 17–26 (2010).
10. Mans, R., Daran, J. G. & Pronk, J. T. Under pressure: evolutionary engineering of yeast strains for improved performance in fuels and chemicals production. *Curr. Opin. Biotechnol.* **50**, 47–56 (2018).
11. Deparis, Q., Claes, A., Foulquie-Moreno, M. R. & Thevelein, J. M. Engineering tolerance to industrially relevant stress factors in yeast cell factories. *FEMS Yeast Res.* **17**, https://doi.org/10.1093/femsyr/fox036 (2017).
12. Caspeta, Y. et al. Altered sterol composition renders yeast thermotolerant. *Science* **346**, 75–78 (2014).
13. Gibney, P. A., Lu, C., Caudy, A. A., Hess, D. C. & Botstein, D. Yeast metabolic and signaling genes are required for heat-shock survival and have little overlap with the heat-induced genes. *Proc. Natl. Acad. Sci. USA* **110**, E4393–E4402 (2013).
14. Lahtvee, P.-J., Kumar, R., Hallström, B. M. & Nielsen, J. Adaptation to different types of stress converge on mitochondrial metabolism. *Mol. Biol. Cell* **27**, 2505–2514 (2016).
15. Firdaus-Raih, M. et al. The *Glaciozyma antarctica* genome reveals an array of systems that provide sustained responses towards temperature variations in a persistently cold habitat. *PLoS ONE* **13**, e0189947 (2018).
16. Silva, A. et al. Regulation of transcription elongation in response to osmostress. *PLoS Genet.* **13**, e1007090 (2017).
17. Hughes Hallett, J. E., Luo, X. & Capaldi, A. P. State transitions in the TORC1 signaling pathway and information processing in *Saccharomyces cerevisiae*. *Genetics* **198**, 773–786 (2014).
18. Kasavi, C., Eraslan, S., Oner, E. T. & Kirdar, B. An integrative analysis of transcriptomic response of ethanol tolerant strains to ethanol in *Saccharomyces cerevisiae*. *Mol. Biosyst.* **12**, 464–476 (2016).
19. Botstein, D. & Fink, G. R. Yeast: an experimental organism for 21st Century biology. *Genetics* **189**, 695–704 (2011).
20. Conesa, A. et al. Blast2GO: a universal tool for annotation, visualization and analysis in functional genomics research. *Bioinformatics* **21**, 3674–3676 (2005).
21. Shen, X. X. et al. Tempo and mode of genome evolution in the budding yeast subphylum. *Cell* **175**, 1533–1545 (2018).
22. Buzzini, P., Turchetti, B. & Yurkov, A. Extremophilic yeasts: the toughest yeasts around? *Yeast* **35**, 487–497 (2018).
23. O'Duibhir, E. et al. Cell cycle population effects in perturbation studies. *Mol. Syst. Biol.* **10**, 732 (2014).
24. Emms, D. M. & Kelly, S. OrthoFinder: solving fundamental biases in whole genome comparisons dramatically improves orthogroup inference accuracy. *Genome Biol.* **16**, 157 (2015).
25. Prachumwat, A. & Li, W.-H. Protein function, connectivity, and duplicability in yeast. *Mol. Biol. Evol.* **23**, 30–39 (2005).
26. Domazet-Lošo, T., Brajković, J. & Tautz, D. A phylostratigraphy approach to uncover the genomic history of major adaptations in metazoan lineages. *Trends Genet.* **23**, 533–539 (2007).
27. Byrne, K. P. & Wolfe, K. H. The yeast gene order browser: combining curated homology and syntenic context reveals gene fate in polyploid species. *Genome Res.* **15**, 1456–1461 (2005).
28. Lahtvee, P.-J. et al. Absolute quantification of protein and mRNA abundances demonstrate variability in gene-specific translation efficiency in yeast. *Cell Syst.* **4**, 495–504.e5 (2017).
29. Wolfe, K. H. Origin of the yeast whole-genome duplication. *PLoS Biol.* **13**, e1002221–e1002221 (2015).
30. Magnan, C. et al. Sequence assembly of *Yarrowia lipolytica* strain W29/CLIB89 shows transposable element diversity. *PLoS ONE* **11**, e0162363 (2016).
31. Giaever, G. et al. Functional profiling of the *Saccharomyces cerevisiae* genome. *Nature* **418**, 387–391 (2002).
32. Pál, C., Papp, B. & Hurst, L. D. Highly expressed genes in yeast evolve slowly. *Genetics* **158**, 927–931 (2001).
33. Mata Bahler, J. J. Correlations between gene expression and gene conservation in fission yeast. *Genome Res.* **13**, 2686–2690 (2003).
34. Huang, C.-J., Lu, M.-Y., Chang, Y.-W. & Li, W.-H. Experimental evolution of yeast for high-temperature tolerance. *Mol. Biol. Evol.* **35**, 1823–1839 (2018).
35. Caspeta, L., Chen, Y. & Nielsen, J. Thermotolerant yeasts selected by adaptive evolution express heat stress response at 30 °C. *Sci. Rep.* **6**, 27003 (2016).
36. Khalturin, K., Hemmrich, G., Fraune, S., Augustin, R. & Bosch, T. C. More than just orphans: are taxonomically-restricted genes important in evolution? *Trends Genet.* **25**, 404–413 (2009).
37. Santos, M. E., Le Bouquin, A., Crumière, A. J. J. & Khila, A. Taxon-restricted genes at the origin of a novel trait allowing access to a new environment. *Science* **358**, 386–390 (2017).
38. Sayah, D. M., Sokolskaja, E., Berthoux, L. & Luban, J. Cyclophilin A retrotransposition into TRIM5 explains owl monkey resistance to HIV-1. *Nature* **430**, 569–573 (2004).
39. Stremlau, M. et al. The cytoplasmic body component TRIM5alpha restricts HIV-1 infection in Old World monkeys. *Nature* **427**, 848–853 (2004).
40. Zhuang, X., Yang, C., Murphy, K. R. & Cheng, C.-H. C. Molecular mechanism and history of non-sense to sense evolution of antifreeze glycoprotein gene in northern gadids. *Proc. Natl. Acad. Sci. USA* **116**, 4400–4405 (2019).
41. Chen, L., DeVries, A. L. & Cheng, C. H. Convergent evolution of antifreeze glycoproteins in Antarctic notothenioid fish and Arctic cod. *Proc. Natl. Acad. Sci. USA* **94**, 3817–3822 (1997).
42. Chen, S., Krinsky, B. H. & Long, M. New genes as drivers of phenotypic evolution. *Nat. Rev. Genet.* **14**, 645–660 (2013).
43. Baalsrud, H. T. et al. De novo gene evolution of antifreeze glycoproteins in codfishes revealed by whole genome sequence data. *Mol. Biol. Evol.* **35**, 593–606 (2017).
44. Hashim, N. H. F. et al. Characterization of Afp1, an antifreeze protein from the psychrophilic yeast Glaciozymaantarctica PI12. *Extremophiles* **17**, 63–73 (2013).
45. Steensma, H. Y. M., de, J. F. C. & Linnekamp, M. The use of electrophoretic karyotypes in the classification of yeasts: *Kluyveromyces marxianus* and *K. lactis*. *Curr. Genet.* **14**, 311–317 (1988).
46. Verduyn, C., Postma, E., Scheffers, W. A. & Van Dijken, J. P. Effect of benzoic acid on metabolic fluxes in yeasts: a continuous-culture study on the regulation of respiration and alcoholic fermentation. *Yeast* **8**, 501–517 (1992).
47. Kim, D., Langmead, B. & Salzberg, S. L. HISAT: a fast spliced aligner with low memory requirements. *Nat. Methods* **12**, 357 (2015).
48. Pertea, M. et al. StringTie enables improved reconstruction of a transcriptome from RNA-seq reads. *Nat. Biotechnol.* **33**, 290 (2015).
49. Dobin, A. et al. STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* **29**, 15–21 (2013).

50. Liao, Y., Smyth, G. K. & Shi, W. featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. *Bioinformatics* **30**, 923–930 (2014).
51. Ritchie, M. E. et al. *limma* powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Res.* **43**, e47 (2015).
52. McCarthy, D. J., Chen, Y. & Smyth, G. K. Differential expression analysis of multifactor RNA-Seq experiments with respect to biological variation. *Nucleic Acids Res.* **40**, 4288–4297 (2012).
53. Wickham, H. et al. Welcome to Tidyverse. *J. Open Source Softw.* **43**, 1–6 (2019).

Acknowledgements

We are grateful to Erik de Hulster and Fredrik Schubert for technical support and training, as well as Gang Li, Lucy Chao, Benjamín Sánchez, and colleagues in the CHASSY consortium for scientific discussions. This project has received funding from the European Union's Horizon 2020 Framework Programme for Research and Innovation—Grant Agreement No. 720824. This work was also supported by the Knut and Alice Wallenberg Foundation and the Novo Nordisk Foundation (Grant no NNF10CC1016517).

Author contributions

T.W.D., N.M., P.A.G., and R.P. performed chemostat cultivations. A.M.-O. performed protein extraction/identification/quantification. I.D. created the analysis pipeline in R and performed differential expression data analysis. T.W.D. prepared/mapped RNA-seq, performed gene grouping analyses, and wrote the manuscript. J.N., C.H., J.-M.D., V.S., and J.M. conceived and supervised the project.

Conflict of interest

The authors declare that they have no conflict of interest.

Additional information

Supplementary information is available for this paper at https://doi.org/10.1038/s41467-020-16073-3.

Correspondence and requests for materials should be addressed to V.S. or J.P.M.

Peer review information *Nature Communications* thanks Rui Alves, and the other, anonymous, reviewer(s) for their contribution to the peer review of this work. Peer reviewer reports are available.

Reprints and permission information is available at http://www.nature.com/reprints

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Paper II:

Evaluating accessibility, usability and interoperability of genome-scale metabolic models for diverse yeasts species

Domenzain, I., Li, F., Kerkhoven, E. J., & Siewers, V.

FEMS Yeast Research, 2021

FEMS Yeast Research, 21, 2021, foab002

doi: 10.1093/femsyr/foab002
Advance Access Publication Date: 11 January 2021
Minireview

MINIREVIEW

Evaluating accessibility, usability and interoperability of genome-scale metabolic models for diverse yeasts species

Iván Domenzain[1,2,†,‡], Feiran Li[1,2,†,§], Eduard J. Kerkhoven[1,2] and Verena Siewers[1,2,*]

[1]Department of Biology and Biological Engineering, Chalmers University of Technology, Kemivägen 10, SE-412 96, Gothenburg, Sweden and [2]Novo Nordisk Foundation Center for Biosustainability, Chalmers University of Technology, Kemivägen 10, SE-412 96, Gothenburg, Sweden

[*]**Corresponding author:** Department of Biology and Biological Engineering, Chalmers University of Technology, Kemivägen 10, SE-412 96 Gothenburg, Sweden. E-mail: siewers@chalmers.se

One sentence summary: Review on computational models of metabolism for diverse yeast species, their development history, applications and critical assessment of their accessibility, usability and interoperability.

[†]Both authors contributed equally to this work.

Editor: Jean-Marc Daran

[‡]Iván Domenzain, http://orcid.org/0000-0002-5322-2040
[§]Feiran Li, http://orcid.org/0000-0001-9155-5260

ABSTRACT

Metabolic network reconstructions have become an important tool for probing cellular metabolism in the field of systems biology. They are used as tools for quantitative prediction but also as scaffolds for further knowledge contextualization. The yeast *Saccharomyces cerevisiae* was one of the first organisms for which a genome-scale metabolic model (GEM) was reconstructed, in 2003, and since then 45 metabolic models have been developed for a wide variety of relevant yeasts species. A systematic evaluation of these models revealed that—despite this long modeling history—the sequential process of tracing model files, setting them up for basic simulation purposes and comparing them across species and even different versions, is still not a generalizable task. These findings call the yeast modeling community to comply to standard practices on model development and sharing in order to make GEMs accessible and useful for a wider public.

Keywords: genome-scale metabolic models; yeast species; systems biology; accessibility; usability; interoperability

INTRODUCTION

Genome-scale metabolic model reconstruction has been established as one of the major modeling approaches for systems-level metabolic studies (Gu *et al.* 2019). These models are mainly built in a bottom-up approach, in which genome information is combined with the accumulated knowledge about the metabolic capabilities of a living organism to reconstruct a complete metabolic map (Nielsen 2017). Another widely used approach for model reconstruction consists of the use of one or multiple well-curated networks as scaffolds, due to the high degree of conservation of metabolism for phylogenetically close species. Metabolic models have been proven to be useful as knowledge databases (Herrgård *et al.* 2008), tools for contextualization of

Received: 14 September 2020; **Accepted:** 8 January 2021

omics data (Kerkhoven *et al.* 2016) and for guiding metabolic engineering projects (Meadows *et al.* 2016), enabling systematic explorations of the relationship between genotypes and phenotypes.

The metabolic model iFF708 (Förster *et al.* 2003) of *Saccharomyces cerevisiae*, the genome of which was the first eukaryotic one to be sequenced (Goffeau *et al.* 1996), was the first published GEM for its entire domain in 2003. This model has been used as a scaffold for further network refinements (Duarte, Herrgård and Palsson 2004; Kuepfer, Sauer and Blank 2005; Herrgård *et al.* 2006; Nookaew *et al.* 2008), which has facilitated the development of metabolic models for several other budding yeast species over the years, due to their well evolutionarily-conserved metabolic capabilities (Shen *et al.* 2018).

Multiple model reconstructions exist, not just for *S. cerevisiae*, but for several other yeast species. These reconstructions have usually been carried out by different research groups, resulting in specific network improvements according to their scientific interests, but at the same time yielding incompatible identifiers for reactions and metabolites hampering any systematic comparison and evaluation across models (Herrgård *et al.* 2008).

As GEMs are valuable tools for a wide variety of applications, their end users vary from academic researchers with different backgrounds and levels of computational skills, to professionals in the biotechnology and pharmaceutical industries. Therefore, there is a strong need for computational metabolic models to be accessible and published in a ready-to-use format, which facilitates their utilization by non-expert users. Additionally, the use of consistent and standardized identifiers for their components enables comparisons across models, thus simplifying the process of finding the best model for a given application.

Latest developments on yeasts GEMs

The development, interconnections and applications of metabolic models for different yeast species have been reviewed extensively (Sánchez and Nielsen 2015; Lopes and Rocha 2017; Castillo, Patil and Jouhten 2019; Chen, Li and Nielsen 2019) however, the list of yeast GEM models is continuously increasing both in number of GEMs and encompassed species. Here we briefly summarize the development history of all models for diverse yeast species that are currently available in the scientific literature. The validation strategies and main applications of these models, as described in their original publications, are provided in Table S3 (Supporting Information), indicating the type of biological data and computational methods used for each case.

S. cerevisiae is one of the most studied organisms in the Eukarya domain, which has resulted in a long modeling history with 18 networks currently available. The models iND750 (Duarte, Herrgård and Palsson 2004), iLL672 (Kuepfer, Sauer and Blank 2005) and iIN800 (Nookaew *et al.* 2008) were directly derived from iFF708 (Förster *et al.* 2003) and subsequently used as templates for iMH805/775 (Herrgård *et al.* 2006), iMM904 (Mo, Palsson and Herrgård 2009), iAZ900 (Zomorrodi and Maranas 2010) and iTO977 (Österlund *et al.* 2013) reconstructions.

As these multiple reconstructions added new knowledge and gap-fills to the network, a first attempt of unification was carried out by the knowledge base Yeast1, published in 2008 (Herrgård *et al.* 2008). The concept of standardized identifiers for reactions and metabolites was first implemented in this reconstruction, but simulation capabilities were not achieved. Sequential curation iterations were performed (Yeast2 and Yeast3) until the publication of Yeast4, which notably increased the network connectivity and the number of included metabolites, making it a suitable model for simulation purposes (Dobson *et al.* 2010). Further updates to the consensus metabolic network have shown to improve predictions on gene essentiality, induced auxotroph phenotypes and cellular growth on diverse environments (Yeast5 (Heavner *et al.* 2012), Yeast6 (Heavner *et al.* 2013) and Yeast7 (Aung, Henry and Walker 2013)). In 2019, a new version of the consensus metabolic network, Yeast8, was published (Lu *et al.* 2019), its reconstruction process combined information from previous GEMs, different curated databases such as KEGG (Kanehisa *et al.* 2016), SGD (Hellerstedt *et al.* 2017), BioCyc (Karp *et al.* 2019), Reactome (Fabregat *et al.* 2018) and UniProt (The UniProt Consortium 2017) and experimental data on substrate usage. Furthermore, Yeast8 provides an ecosystem of multilayer models suited for different kinds of phenotype predictions, ranging from 1011 strain-specific models to incorporation of enzyme constraints (ecYeast8) and protein 3D structures (proYeast; Lu *et al.* 2019).

In parallel with the development of the consensus network, iSce926 (Chowdhury, Chowdhury and Maranas 2015) was derived from Yeast7 (Aung, Henry and Walker 2013) in 2015, incorporating gene essentiality and synthetic lethality information to curate gene-reaction rules. The model iSc-AMRS-1 (Wichmann *et al.* 2016) was developed from iLL672 (Kuepfer, Sauer and Blank 2005) in 2016, mainly by curation of proton balancing for mitochondrial ATP production and reaction reversibility, aiming to improve flux distribution predictions in order to investigate production of isopropenoids.

The model SpoMBEL1693 for *Schizosaccharomyces pombe*, a model organism for eukaryotic cell cycle studies, was developed in 2012 using annotated genes and reactions from the KEGG database as a draft network (Sohn *et al.* 2012). iNX804, a metabolic model for *Candida glabrata*, known as a platform organism for pyruvate production, was reconstructed in 2013 and used for identification of gene targets for enhanced production of pyruvate-derived fine chemicals (Xu *et al.* 2013). The metabolism of *Candida tropicalis*, known as a promising host for α, ω-dicarboxylic acids production, has been studied with the model iCT646, reconstructed through the collection of multiple database information in 2016 (Mishra *et al.* 2016). The model iOD907, a metabolic network for *Kluyveromyces lactis*, a yeast commonly used in the dairy industry, was published in 2014 (Dias *et al.* 2014). Its reconstruction process used iMM904, for *S. cerevisiae*, as a scaffold and merged it with annotation for metabolic genes and transporters from KEGG (Kanehisa *et al.* 2017) and TCDB (Saier *et al.* 2016), respectively. This model was validated with data for growth on diverse carbon sources and used to investigate phenotypic differences for single gene knockout strains between *K. lactis* and *S. cerevisiae* (Dias *et al.* 2014).

Pichia pastoris is an established workhorse in biotechnology for heterologous protein production, as it shows superior protein secretion efficiency compared with other yeasts (Schmidt 2004). Additionally, humanized N-glycosylation patterns for recombinant protein production can be obtained by engineering its metabolism. The first two GEMs for *P. pastoris*, PpaMBEL1254 (Sohn *et al.* 2010) and iPP668 (Tomàs-Gamisans, Ferrer and Albiol 2016), were both developed in 2010 using genome annotation information from databases and literature. In 2015, ihGlycopastoris (Irani *et al.* 2016) was specially developed for simulation of recombinant protein production as a target, by combining the previously established iLC915 (Caspeta *et al.* 2012) model with humanized N-glycosylation pathways. This allowed the investigation of the influence of N-glycosylation processes on protein

production and the model was used for the prediction of gene overexpression targets for improving protein yields. The model Kp.1.0 was published in 2017, in which 12 different biomass compositions were tested under different growth conditions, showing minor effects on growth and gene essentiality predictions, but drastic changes in flux distributions (Cankorur-Cetinkaya, Dikicioglu and Oliver 2017). A total of three previous *P. pastoris* reconstructions (Chung *et al.* 2010; Sohn *et al.* 2010; Caspeta *et al.* 2012) were merged into iMT1026 (Tomàs-Gamisans, Ferrer and Albiol 2016), expanding the representation of fatty acid and sphingolipid metabolism, intact N-glycosylation, O-glycosylation and glycosylphosphatidylinositol(GPI)-anchor pathways. iMT1026 was then curated to iMT1026.v3 in 2018, leading to a refinement of predictions for cellular growth on glycerol and methanol as carbon sources (Tomàs-Gamisans, Ferrer and Albiol 2018). Additionally, the model iRY1243 was created in 2017 by merging iPP668, PpaMBEL1254, iLC915 and iMT1026, also incorporating curation of biosynthesis of vitamins and cofactors, which added more than 200 metabolic genes to the network. This model was validated with the use of RNAseq data for different conditions, utilization of carbon and nitrogen sources and ^{13}C-labeled derived fluxomics, yielding an overall high consistency of predictions for essential genes, flux distributions and different mutant phenotypes (Ye *et al.* 2017).

The yeast *Scheffersomyces stipitis* (formerly known as *Pichia stipitis*) has raised interest due to its great native potential for xylose utilization. In 2012, three models were published for this species: iTL885 (Liu *et al.* 2012) and iSS884 (Caspeta *et al.* 2012) were derived from previous *S. cerevisiae*'s models, whilst iBB814 (Balagurunathan *et al.* 2012) was reconstructed from genome annotation extracted from various databases. A modified version of iBB814, the model iDH814, was published in 2016 and used to elucidate the redox balance shift response to reduced oxygen supply conditions (Hilliard *et al.* 2018). As these four reconstructions just account for the cytoplasm, mitochondria and peroxisome as cellular compartments, a fully compartmentalized model for this relevant organism is still missing.

The oleaginous yeast *Yarrowia lipolytica*, is another organism for which multiple GEMs already exist. Its first model, iNL895 developed in 2012 (Loira *et al.* 2012) and other two following models iMK735 (Kavščcek *et al.* 2015) and iYali4 (Kerkhoven *et al.* 2016), were derived from previous networks of the phylogenetically distant yeast *S. cerevisiae*, in contrast to iYL619_PCP (Pan and Hua 2012), reconstructed directly from *Y. lipolytica* specific information available in public databases and literature. In 2018, iYLI647 (Mishra *et al.* 2018) was developed using a previous reconstruction for the same species, iMK735 (Kavščcek *et al.* 2015), as a scaffold and expanded to include the ω-oxidation pathway that converts fatty acids to long-chain dicarboxylic acids (DCAs), the subsequent fatty-acid degrading β-oxidation pathway and branched-chain amino acid degradation pathways, in order to guide simulation of metabolic engineering strategies for enhanced DCA production.

During these years, other non-conventional yeasts have gained more attention due to their fascinating and diverse phenotypes. Several GEMs have been constructed as an attempt to understand their particular traits. *Rhodotorula toruloides* is an oleaginous yeast, which can accumulate lipids up to 70% of its dry mass (Ratledge and Wynn 2002). Previous modeling approaches have explored the use of constraint-based methods together with a reduced metabolic network for this organism to assess lipid accumulation on different substrates (Bommareddy *et al.* 2015; Castañeda *et al.* 2018), but its first genome-scale model, rthoGEM (Tiukova *et al.* 2019), was published in 2019. Cell growth data using glucose, xylose and glycerol as substrates were used to validate the model, while gene targets for triacylglycerol and carotenoid production were predicted with the use of the FSEOF algorithm (Choi *et al.* 2010). That same year, iRhto1108 (Dinh *et al.* 2019), was developed using Yeast7 and the Kbase fungal metabolic network (Arkin *et al.* 2018) as model templates. This model increased the metabolic gene coverage in comparison to rthoGEM (from 926 to 1108) and enabled growth simulations using arabinose and cellobiose as carbon sources.

Zygosaccharomyces bailii has been described to have high tolerance towards acetic acid (Palma *et al.* 2017; Palma, Guerreiro and Sá-Correia 2018). It has been suggested that the *Zygosaccharomyces* clade diverged from *Saccharomyces* ancestors just before the whole genome duplication event (WGD; Kurtzman 2003), which took place approximately 100 million years ago, making the *Zygosaccharomyces* genus the closest pre-WGD ancestral group of relatives to study the genome evolution of *S. cerevisiae* (Hagman *et al.* 2013; Solieri *et al.* 2013). The model ZyPa1 (Filippo *et al.* 2018) was reconstructed using homology information from 20 different yeasts belonging to the Saccharomycetaceae family, and was then connected to the KEGG database to obtain a draft network. Stoichiometry and localization information for the reactions were extracted from the models Yeast7 (Aung, Henry and Walker 2013) and iOD907 (Dias *et al.* 2014). ZyPa1 contains 2413 genes, more than twice the number of genes in Yeast8 (Lu *et al.* 2019), being the metabolic model for a yeast species with the highest number of genes. This GEM has been applied to the study of cellular growth under co-consumption of lactate and glucose.

Kluyveromyces marxianus is a thermotolerant yeast that can even tolerate temperatures as extreme as 52°C (Nonklang *et al.* 2008), making it a specially interesting organism host for industrial bioproduction. The first GEM for *K. marxianus*, iSM996, was built in 2019 (Marcišauskas, Ji and Nielsen 2019) by using a draft model generated with the RAVEN Toolbox (Wang *et al.* 2018), aided by the KEGG database and the models iOD907 (Dias *et al.* 2014) and Yeast7 (Aung, Henry and Walker 2013) as sources for the network gap-filling process. iSM996 was validated using data on carbon and nitrogen source usage, and transcriptome datasets were integrated in order to simulate growth under different temperatures (Marcišauskas, Ji and Nielsen 2019).

Lachancea kluyveri is a weak Crabtree positive yeast of industrial relevance due to its capabilities for ethyl-acetate secretion, when cultivated in aerobic batch conditions, and usage of urea and uracil as sole nitrogen sources for growth. In 2020, the model iPN730 (Ghosh *et al.* 2020) was built on a Kbase workspace (Arkin *et al.* 2018) using iMM904 (Mo, Palsson and Herrgård 2009) for *S. cerevisiae* as a template network and other 13 fungi models as references for homologous reactions searches. The model was validated by simulating cellular growth on diverse environments (Ghosh *et al.* 2020).

A repository for yeast species metabolic models

All aforementioned yeasts GEMs, together with the previously published models, were used to query the literature using the keyword 'yeast' together with 'metabolic model', 'GSM', 'GEM' or 'GENRE' (genome-scale network reconstruction). In total, 43 model files for 12 different organisms were found either as part of publications in peer-reviewed journals, supplementary files for preprint articles in bioRxiv, or in the *yeastnet* model database (https://sourceforge.net/projects/yeast) when no specific publication about their reconstruction was found (as in the case of Yeast2, Yeast3 and Yeast4). Most of these yeast species belong

to the Saccharomycetales order in the Ascomycota phylum, but some of them have been classified as part of other classes, as *Schizosaccharomyces pombe* (Schizosaccharomycetes) or even phyla, such as the Basidiomycota fungus *Rhodotorula toruloides* (Table S1, Supporting Information).

As expected, *S. cerevisiae* is the yeast species for which the most GEMs have been reconstructed, however multiple models are also available for *P. pastoris*, *Y. lipolytica* and *S. stipitis* (Fig. 1A). This collection of model files has been stored in a publicly available GitHub repository at https://github.com/SysBioChalmers/YeastsModels, together with the necessary scripts for their further analysis. The search and exploration processes for these models pointed out several aspects that can be classified into three main categories: accessibility, usability and interoperability.

Model accessibility

The analyzed models in this review span more than 17 years of active research, in which standards for file formats and sharing practices in the field of systems biology have changed, making the retrieval of their original files a time-consuming and not automatable task. Even though the Systems Biology Markup Language (SBML) was released in 2002 (Hucka *et al.* 2003), and since then has evolved to become the standard file format for metabolic modeling, 27% of the analyzed models were shared in a different format in their original publications, such as .txt, .XLS and .pdf (Fig. 1B and C), which limits scientific exchange and reproducibility of results on different setups due to their dependence on specific software applications (Ravikrishnan and Raman 2015).

As not all models could be successfully obtained from their original sources, models were also sought in other public repositories such as Biomodels (Chelliah *et al.* 2015), Biomet (Garcia-Albornoz *et al.* 2014) and openCOBRA models (Ebrahim *et al.* 2015; Fig. 1D), which contain curated metabolic reconstructions not just for yeast species but for all key phylogenetic groups (Monk, Nogales and Palsson 2014). The models from the last decade present in this catalogue reflect the trend of referring to unambiguous entries in such databases instead of uploading model files as supplementary material to their respective journal websites.

Notably, a novel methodology for model sharing and development has been proposed by the Yeast8 project (Lu *et al.* 2019) and the Memote model test suite (Lieven *et al.* 2020), which with the aid of version control tools, such as Git and GitHub, provides not just the final snapshot of a GEM but its whole development history, offering also a web platform for open and continuous development. These version control tools have also been implemented for *Y. lipolytica*, *K. marxianus* and *R. toruloides* GEMs (iYali4 (Kerkhoven *et al.* 2016), iSM996 (Marcišauskas, Ji and Nielsen 2019), rthoGEM (Tiukova *et al.* 2019) and iRhto1108 (Dinh *et al.* 2019)), which represent 11% of the collected models (Fig. 1E). More community-driven modeling efforts are expected to emerge in the next years as a way to circumvent the drawback of having multiple independent reconstructions available for some of these yeast species.

Model usability

In order to evaluate the complexity of the process of getting started when utilizing a GEM, a testing pipeline was developed using the RAVEN (Wang *et al.* 2018), COBRA (Heirendt *et al.* 2019) and COBRApy (Ebrahim *et al.* 2013) toolboxes, which in a series of sequential steps aims to obtain feasible flux balance analysis simulations (Orth, Thiele and Palsson 2010), with cellular growth maximization as an objective function, assuming that no prior knowledge about the model´s specific structure and identifiers was available. In total, SBML files for 37 models were found available in this study, and therefore analyzed by the mentioned pipeline.

The first tested functionality was the importability of each SBML model into a non-empty MATLAB structure (Table S2, Supporting Information). This was satisfactorily achieved for the majority of these models, 97%. The only non-loadable SBML file was also tested with the COBRApy toolbox, but its import could not be accomplished due to parsing errors. Secondly, a default objective function was sought in the model structure by retrieving any non-zero coefficient in the objective function field or so called 'c vector'. Of the analyzed models, 76% showed a predefined objective function. Further exploration found that all of these objectives are maximization of the growth rate, 'biomass exchange' or 'biomass formation'. Taking this into account, traceability of a biomass pseudoreaction was also evaluated. For doing so, the presence of the substrings 'growth', 'biomass' and 'vgro' was explored in the model.rxns and model.rxnNames fields. In total 84% of the tested models contain a biomass pseudoreaction identifiable with the used patterns. This does not imply that a biomass reaction is absent for the 16% remaining models, but that the search for it would require a customized manual procedure for each of them.

For all of these models, maximization of the found biomass reaction was set as an objective function and all of their exchange reactions were opened in both directions (lower and upper bounds of −1000 and 1000 mmol/gDw h, respectively) to check *in silico* cellular growth capabilities. In total, 76% of the tested subset (28 models) showed a non-zero growth rate when subject to these constraints. We consider these models as available in a ready-to-use setup, as no further steps or manual inspection was needed to simulate growth. Detailed information for the evaluated metrics and features can be found in Table S2 (Supporting Information).

In order to assess the utilization of these models by the scientific community, the total and average annual citations were used as proxy metrics. Figure 2E shows that a larger proportion of the cited models that were recently published (<5 years ago) have been made available in a ready-to-use format (77%) in comparison to those that were published a longer time ago (62%). For the *S. cerevisiae* network reconstructions, it is clear that older models are on average more used or referred to in the scientific literature. However, as time has passed more models have become available and decays on citations for older models usually coincide with publication and rise of newer ones (Fig. 2F). This might suggest that scientific interest shifts towards more recent models as they accumulate the knowledge gathered by previous reconstruction iterations.

Interoperability

As described above and repeatedly concluded (Dräger and Palsson 2014; Ebrahim *et al.* 2015; Heavner and Price 2015; Sánchez and Nielsen 2015; Mendoza *et al.* 2019), the lack of identifier consistency and connection to external databases for all of the relevant components of GEMs (metabolites, reactions, genes and cellular compartments) together with the use of non-standardized file formats, are the main obstacles for direct model comparison and assessment, even across reconstructions for a single species.

(A) Models per organism

Species	Models
Saccharomyces cerevisiae	19
Komagataella pastoris	8
Yarrowia lipolytica	5
Scheffersomyces stipitis	4
Rhodotorula toruloides	2
Candida glabrata	1
Kluyveromyces lactis	1
Zygosaccharomyces parabilii	1
Kluyveromyces marxianus	1
Candida tropicalis	1
Schizosaccharomyces pombe	1
Lachancea kluyveri	1

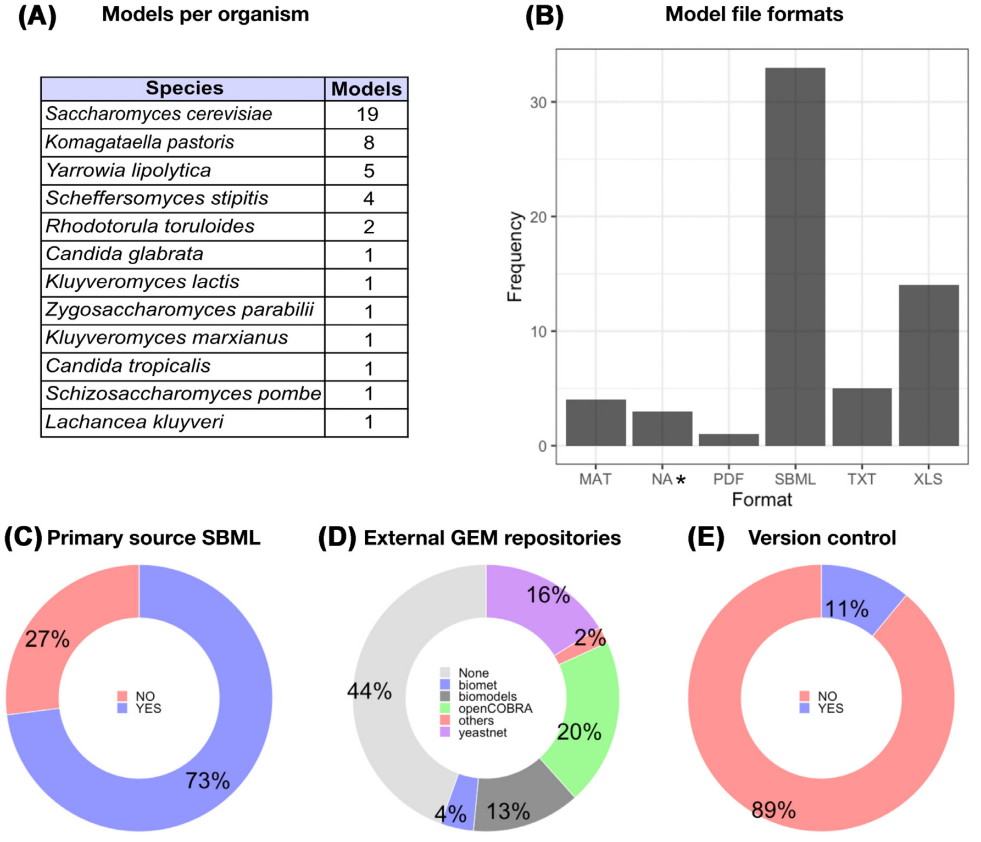


Figure 1. Accessibility of metabolic models for diverse yeast species. **(A)** Number of published models per species. **(B)** Number of published models per file format. Models available in several formats are counted multiple times. *NA indicates model files that were not available in either their original publications or external model repositories **(C)** Proportion of models provided as an SBML file in their original source or publication. **(D)** Proportion of yeast models stored in different public databases. Models stored in several databases are just accounted as part of the one that uploaded them first. **(E)** Proportion of models with continuous development tracked on public repositories.

In order to aid systematic model development, according to community-agreed practices, a standardized set of metabolic model tests (Memote) has recently been developed as an open-source software suite (Lieven *et al.* 2020). Memote tests are divided into organism- and model-specific ones, not applicable to all reconstructions, and a section of independent tests, which check for model consistency (in terms of mass and charge balance, metabolite connectivity and stoichiometric consistency), and annotation, or connection to external databases, for metabolites, reactions, genes and SBO terms (systems biology ontology terms; Courtot *et al.* 2011). This pipeline assigns a numerical score, based on the specific model characteristics, to each of the independent tests, relevant for comparing evolution of particular model features across versions.

The 37 SBML model files analyzed above were furthermore tested by the Memote suite. As this software relies on the latest version of the SBML Level 3 Flux Balance Constraints package (Olivier and Bergmann 2018), not all of the models could be tested due to parsing errors for those available in previous or conflicting SBML versions (36%), as shown in Fig. 3A. Noteworthy, this is not an indicator of model quality or predictive performance, but rather one of compliance with model format standards. Further details for all of the individual tests and computed scores are available as HTML reports and also as part of Table S2 (Supporting Information), both stored in the aforementioned GitHub repository.

The community-driven series of consensus metabolic network reconstructions for *S. cerevisiae* has tried to overcome some of the obstacles mentioned above by keeping consistency of identifiers across the subsequent model refinement iterations. However, this approach has not yet been applied to any of the other yeast species models analyzed in this review. Such consistency allows to interpret Memote standardized test results as an evolution of the network in different regards, offering a systematic guidance for further development. Annotation of metabolites, reactions and SBO terms has been improved throughout the different versions of the *S. cerevisiae* model (Fig. 3B). Resultingly, Yeast8 shows the most complete degree of annotation for all of these features, even though standardized gene identifiers that are traceable to an external database are still missing.

CONCLUSIONS

Here we reviewed, collected and evaluated the usability of the available GEMs for different yeasts species, offering a valuable concentrated resource for the community. The model recollection process evidenced that not all of them are easily accessible and multiple sources were needed to be queried. Even though specialized databases for curated GEMs exist, connections between them are still missing, which might hamper large-scale multi-species studies. We also found that GEM files have been shared in a wide variety of file formats, making the uti-

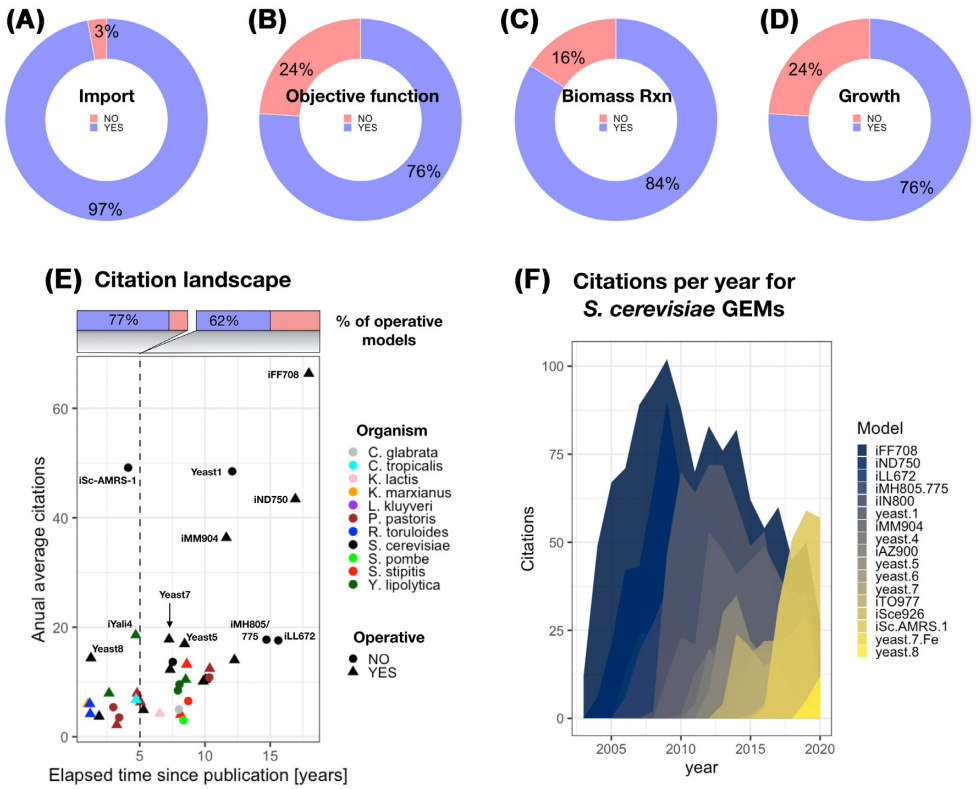


Figure 2. Model usability. **(A)** Proportion of tested SBML models successfully imported with the RAVEN, COBRA or COBRApy toolboxes (total = 37 models). **(B)** Proportion of tested models with a default objective function. **(C)** Proportion of tested models with a biomass pseudoreaction identifiable with the substrings 'biomass', 'growth' or 'vgro'. **(D)** Proportion of models yielding a non-zero growth rate according to the developed testing pipeline. **(E)** Citation landscape of models of yeasts metabolism. Annual average citations vs elapsed time since publication per species, the proportion of 'operative models' (available in a ready-to-use format, according to the developed testing pipeline) is indicated in the upper part for models that have been published more or less than 5 years ago. **(F)** Evolution of the annual citations for models of *S. cerevisiae* metabolism. Citations were queried from Google scholar, accessed on September 4th, 2020.

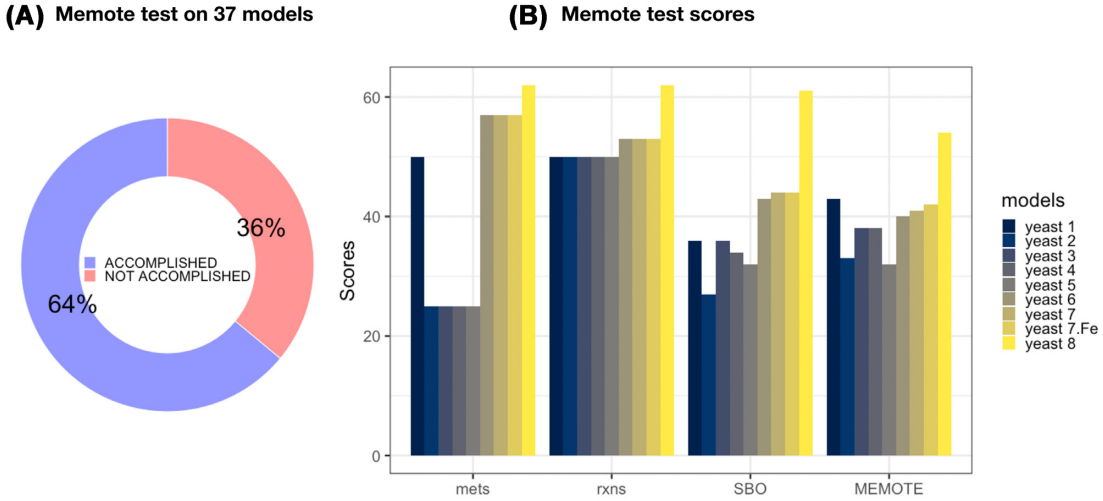


Figure 3. Memote tests results. **(A)** Proportion of models for which the automated Memote test was accomplished. **(B)** Memote test scores for the consensus reconstructions of the *S. cerevisiae* metabolic network. Scores for metabolites, reactions and SBO terms evaluate the degree of annotation for such components with external databases identifiers that can facilitate the traceability of a component across different model versions. The Memote global score takes into account the structure, consistency, annotation and functionality of metabolic models.

lization of some of them dependent on specific software tools. Storing and sharing models using the latest version of the standard SBML format will facilitate scientific exchange and enable reproducibility of results, avoiding platform dependent parsing issues.

As part of this review, a simplified model test pipeline was developed and run for all of the yeast GEMs with an available SBML file. With the aim of obtaining feasible FBA simulations with the minimal number of steps, we simulated the initial familiarization process of a non-expert user with a new model.

It was found that 28 of the tested models (representing 62% of the models in this catalogue) were available in a ready-to-use format, as *in-silico* growth was obtained without any further knowledge or utilization experience on them. This result must not be interpreted as a measurement of model quality, as biological meaningfulness or consistency of predictions were not evaluated. More robust tests were performed with the aid of the Memote suite. Nonetheless, this was not possible for all of the analyzed models due to outdated file formats. For such cases, update of their respective SBML files is recommended in order to ensure compatibility with the latest modeling and analysis tools and to facilitate further development. The results of the Memote standardized tests illustrated a progressive evolution concerning the annotation of model components for the different versions of the *S. cerevisiae* metabolic network, highlighting the advantages of community-driven model development.

The total or partial lack of cross-references of model components to widely used external databases is still a common trait of the models in this catalogue. GEMs are usually described as valuable scientific resources not just for quantitative predictions but as genome-scale knowledgebases of living organisms. However, as their usability and exploration are still hindered by the lack of format consistency, cross-references and continuous community development, the full exploitation of their potential remains restricted to expert users.

ACKNOWLEDGMENTS

We are grateful to Benjamin D. Heavner for sharing his repository of *S. cerevisiae* models.

SUPPLEMENTARY DATA

Supplementary data are available at *FEMSYR* online.

FUNDING

This work has received funding from the European Union's Horizon 2020 Framework Programme for Research and Innovation—Grant Agreement Number 720 824 [to I.D.] and Grant Agreement Number 686 070 [to F.L.].

Conflicts of Interest. The authors declare that this research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

REFERENCES

Arkin AP, Cottingham RW, Henry CS *et al.* KBase: the United States department of energy systems biology knowledgebase. *Nat Biotechnol* 2018, DOI: 10.1038/nbt.4163.

Aung HW, Henry SA, Walker LP. Revising the representation of fatty acid, glycerolipid, and glycerophospholipid metabolism in the consensus model of yeast metabolism. *Ind Biotechnol* 2013, DOI: 10.1089/ind.2013.0013.

Balagurunathan B, Jonnalagadda S, Tan L *et al.* Reconstruction and analysis of a genome-scale metabolic model for Scheffersomyces stipitis. *Microb Cell Fact* 2012;**11**, DOI: 10.1186/1475-2859-11-27.

Bommareddy RR, Sabra W, Maheshwari G *et al.* Metabolic network analysis and experimental study of lipid production in *Rhodosporidium toruloides* grown on single and mixed substrates. *Microb Cell Fact* 2015, DOI: 10.1186/s12934-015-0217-5.

Cankorur-Cetinkaya A, Dikicioglu D, Oliver SG. Metabolic modeling to identify engineering targets for *Komagataella phaffii*: the effect of biomass composition on gene target identification. *Biotechnol Bioeng* 2017;**114**:2605–15.

Caspeta L, Shoaie S, Agren R *et al.* Genome-scale metabolic reconstructions of *Pichia stipitis* and *Pichia pastoris* and in silico evaluation of their potentials. *BMC Syst Biol* 2012, DOI: 10.1186/1752-0509-6-24.

Castañeda MT, Nuñez S, Garelli F *et al.* Comprehensive analysis of a metabolic model for lipid production in *Rhodosporidium toruloides*. *J Biotechnol* 2018, DOI: 10.1016/j.jbiotec.2018.05.010.

Castillo S, Patil KR, Jouhten P. Yeast genome-scale metabolic models for simulating genotype-phenotype relations. *Prog Mol Subcell Biol* 2019, DOI: 10.1007/978-3-030-13035-0_5.

Chelliah V, Juty N, Ajmera I *et al.* BioModels: ten-year anniversary. *Nucleic Acids Res* 2015, DOI: 10.1093/nar/gku1181.

Chen Y, Li G, Nielsen J. Genome-scale metabolic modeling from yeast to human cell models of complex diseases: latest advances and challenges. *Methods Mol Biol* 2019;**2049**:329–45.

Choi HS, Lee SY, Kim TY *et al.* In silico identification of gene amplification targets for improvement of lycopene production. *Appl Environ Microbiol* 2010;**76**:3097–105.

Chowdhury R, Chowdhury A, Maranas CD. Using gene essentiality and synthetic lethality information to correct yeast and CHO cell genome-scale models. *Metabolites* 2015;**5**:536–70.

Chung BKS, Selvarasu S, Andrea C *et al.* Genome-scale metabolic reconstruction and in silico analysis of methylotrophic yeast *Pichia pastoris* for strain improvement. *Microb Cell Fact* 2010;**9**:1–15.

Courtot M, Juty N, Knüpfer C *et al.* Controlled vocabularies and semantics in systems biology. *Mol Syst Biol* 2011;**7**:543.

Dias O, Pereira R, Gombert AK *et al.* iOD907, the first genome-scale metabolic model for the milk yeast *Kluyveromyces lactis*. *Biotechnol J* 2014, DOI: 10.1002/biot.201300242.

Dinh H V., Suthers PF, Chan SHJ *et al.* A comprehensive genome-scale model for *Rhodosporidium toruloides* IFO0880 accounting for functional genomics and phenotypic data. *Metab Eng Commun* 2019, DOI: 10.1016/j.mec.2019.e00101.

Dobson PD, Smallbone K, Jameson D *et al.* Further developments towards a genome-scale metabolic model of yeast. *BMC Syst Biol* 2010;**4**:145.

Dräger A, Palsson B. Improving collaboration by standardization efforts in systems biology. *Front Bioeng Biotechnol* 2014;**2**, DOI: 10.3389/fbioe.2014.00061.

Duarte NC, Herrgård MJ, Palsson B. Reconstruction and validation of *Saccharomyces cerevisiae* iND750, a fully compartmentalized genome-scale metabolic model. *Genome Res* 2004, DOI: 10.1101/gr.2250904.

Ebrahim A, Almaas E, Bauer E *et al.* Do genome-scale models need exact solvers or clearer standards? *Mol Syst Biol* 2015, DOI: 10.15252/msb.20156157.

Ebrahim A, Lerman JA, Palsson BO *et al.* COBRApy: constraints-based reconstruction and analysis for python. *BMC Syst Biol* 2013;**7**:74.

Fabregat A, Jupe S, Matthews L *et al.* The reactome pathway knowledgebase. *Nucleic Acids Res* 2018;**46**:D649–55.

Filippo M Di, Ortiz-Merino RA, Damiani C *et al.* Genome-scale metabolic reconstruction of the stress-tolerant hybrid yeast *Zygosaccharomyces parabailii*. *bioRxiv* 2018:373621. DOI: 10.1101/373621.

Förster J, Famili I, Fu P *et al.* Genome-scale reconstruction of the *Saccharomyces cerevisiae* metabolic network. *Genome Res* 2003, DOI: 10.1101/gr.234503.

Garcia-Albornoz M, Thankaswamy-Kosalai S, Nilsson A *et al.* BioMet Toolbox 2.0: genome-wide analysis of metabolism and omics data. *Nucleic Acids Res* 2014, DOI: 10.1093/nar/gku371.

Ghosh A, Nanda P, Patra P *et al. Reconstruction and Analysis of Genome-Scale Metabolic Model of Weak Crabtree Positive Yeast Lachancea kluyveri*. 2020. DOI: 10.21203/rs.2.16651/v1.

Goffeau A, Barrell G, Bussey H *et al.* Life with 6000 genes. *Science (80-)* 1996, DOI: 10.1126/science.274.5287.546.

Gu C, Kim GB, Kim WJ *et al.* Current status and applications of genome-scale metabolic models. *Genome Biol* 2019;**20**:121.

Hagman A, Säll T, Compagno C *et al.* Yeast "Make-Accumulate-Consume" life strategy evolved as a multi-step process that predates the whole genome duplication. *PLoS One* 2013, DOI: 10.1371/journal.pone.0068734.

Heavner BD, Price ND. Comparative analysis of yeast metabolic network models highlights progress, opportunities for metabolic reconstruction. *PLoS Comput Biol* 2015, DOI: 10.1371/journal.pcbi.1004530.

Heavner BD, Smallbone K, Barker B *et al.* Yeast 5 - an expanded reconstruction of the *Saccharomyces cerevisiae* metabolic network. *BMC Syst Biol* 2012, DOI: 10.1186/1752-0509-6-55.

Heavner BD, Smallbone K, Price ND *et al.* Version 6 of the consensus yeast metabolic network refines biochemical coverage and improves model performance. *Database* 2013, DOI: 10.1093/database/bat059.

Heirendt L, Arreckx S, Pfau T *et al.* Creation and analysis of biochemical constraint-based models using the COBRA Toolbox v.3.0. *Nat Protoc* 2019;**14**:639–702.

Hellerstedt ST, Nash RS, Weng S *et al.* Curated protein information in the Saccharomyces genome database. *Database (Oxford)* 2017;**2017**:bax011.

Herrgård MJ, Lee BS, Portnoy V *et al.* Integrated analysis of regulatory and metabolic networks reveals novel regulatory mechanisms in *Saccharomyces cerevisiae*. *Genome Res* 2006, DOI: 10.1101/gr.4083206.

Herrgård MJ, Swainston N, Dobson P *et al.* A consensus yeast metabolic network reconstruction obtained from a community approach to systems biology. *Nat Biotechnol* 2008;**26**:1155–60.

Hilliard M, Damiani A, He QP *et al.* Elucidating redox balance shift in *Scheffersomyces stipitis*' fermentative metabolism using a modified genome-scale metabolic model. *Microb Cell Fact* 2018;**17**:140.

Hucka M, Finney A, Sauro HM *et al.* The systems biology markup language (SBML): a medium for representation and exchange of biochemical network models. *Bioinformatics* 2003;**19**:524–31.

Irani ZA, Kerkhoven EJ, Shojaosadati SA *et al.* Genome-scale metabolic model of *Pichia pastoris* with native and humanized glycosylation of recombinant proteins. *Biotechnol Bioeng* 2016;**113**:961–9.

Kanehisa M, Furumichi M, Tanabe M *et al.* KEGG: new perspectives on genomes, pathways, diseases and drugs. *Nucleic Acids Res* 2016;**45**:D353–61.

Kanehisa M, Furumichi M, Tanabe M *et al.* KEGG: new perspectives on genomes, pathways, diseases and drugs. *Nucleic Acids Res* 2017;**45**:D353–61.

Karp PD, Billington R, Caspi R *et al.* The BioCyc collection of microbial genomes and metabolic pathways. *Brief Bioinform* 2019;**20**:1085–93.

Kavšcek M, Bhutada G, Madl T *et al.* Optimization of lipid production with a genome-scale model of *Yarrowia lipolytica*. *BMC Syst Biol* 2015, DOI: 10.1186/s12918-015-0217-4.

Kerkhoven EJ, Pomraning KR, Baker SE *et al.* Regulation of amino-acid metabolism controls flux to lipid accumulation in *Yarrowia lipolytica*. *npj Syst Biol Appl* 2016;**2**:1–7.

Kuepfer L, Sauer U, Blank LM. Metabolic functions of duplicate genes in *Saccharomyces cerevisiae*. *Genome Res* 2005, DOI: 10.1101/gr.3992505.

Kurtzman CP. Phylogenetic circumscription of *Saccharomyces, Kluyveromyces* and other members of the Saccharomycetaceae, and the proposal of the new genera Lachancea, Nakaseomyces, Naumovia, Vanderwaltozyma and Zygotorulaspora. *FEMS Yeast Res* 2003;**4**:233–45.

Lieven C, Beber ME, Olivier BG *et al.* MEMOTE for standardized genome-scale metabolic model testing. *Nat Biotechnol* 2020, DOI: 10.1038/s41587-020-0446-y.

Liu T, Zou W, Liu L *et al.* A constraint-based model of *Scheffersomyces stipitis* for improved ethanol production. *Biotechnol Biofuels* 2012;**5**:72.

Loira N, Dulermo T, Nicaud JM *et al.* A genome-scale metabolic model of the lipid-accumulating yeast *Yarrowia lipolytica*. *BMC Syst Biol* 2012;**6**, DOI: 10.1186/1752-0509-6-35.

Lopes H, Rocha I. Genome-scale modeling of yeast: chronology, applications and critical perspectives. *FEMS Yeast Res* 2017, DOI: 10.1093/femsyr/fox050.

Lu H, Li F, Sánchez BJ *et al.* A consensus *S. cerevisiae* metabolic model Yeast8 and its ecosystem for comprehensively probing cellular metabolism. *Nat Commun* 2019;**10**:1–13.

Marcišauskas S, Ji B, Nielsen J. Reconstruction and analysis of a *Kluyveromyces marxianus* genome-scale metabolic model. *BMC Bioinformatics* 2019;**20**:551.

Meadows AL, Hawkins KM, Tsegaye Y *et al.* Rewriting yeast central carbon metabolism for industrial isoprenoid production. *Nature* 2016, DOI: 10.1038/nature19769.

Mendoza SN, Olivier BG, Molenaar D *et al.* A systematic assessment of current genome-scale metabolic reconstruction tools. *Genome Biol* 2019, DOI: 10.1186/s13059-019-1769-1.

Mishra P, Lee NR, Lakshmanan M *et al.* Genome-scale model-driven strain design for dicarboxylic acid production in *Yarrowia lipolytica*. *BMC Syst Biol* 2018, DOI: 10.1186/s12918-018-0542-5.

Mishra P, Park GY, Lakshmanan M *et al.* Genome-scale metabolic modeling and in silico analysis of lipid accumulating yeast *Candida tropicalis* for dicarboxylic acid production. *Biotechnol Bioeng* 2016, DOI: 10.1002/bit.25955.

Mo ML, Palsson BO, Herrgård MJ. Connecting extracellular metabolomic measurements to intracellular flux states in yeast. *BMC Syst Biol* 2009;**3**:37.

Monk J, Nogales J, Palsson BO. Optimizing genome-scale network reconstructions. *Nat Biotechnol* 2014, DOI: 10.1038/nbt.2870.

Nielsen J. Systems biology of metabolism. *Annu Rev Biochem* 2017;**86**:245–75.

Nonklang S, Abdel-Banat BMA, Cha-aim K *et al.* High-temperature ethanol fermentation and transformation with linear DNA in the thermotolerant yeast *Kluyveromyces marxianus* DMKU3-1042. *Appl Environ Microbiol* 2008, DOI: 10.1128/AEM.01854-08.

Nookaew I, Jewett MC, Meechai A *et al.* The genome-scale metabolic model iIN800 of *Saccharomyces cerevisiae* and its validation: a scaffold to query lipid metabolism. *BMC Syst Biol* 2008, DOI: 10.1186/1752-0509-2-71.

Olivier BG, Bergmann FT. SBML Level 3 package: flux balance constraints version 2 SBML Level 3 package: flux balance constraints ('fbc'). *J Integr Bioinform* 2018, DOI: 10.1515/jib-2017-0082.

Orth JD, Thiele I, Palsson BØ. What is flux balance analysis? *Nat Biotechnol* 2010;**28**:245–8.

Österlund T, Nookaew I, Bordel S *et al*. Mapping condition-dependent regulation of metabolism in yeast through genome-scale modeling. *BMC Syst Biol* 2013;**7**:36.

Palma M, Dias PJ, Roque F de C *et al*. The *Zygosaccharomyces bailii* transcription factor Haa1 is required for acetic acid and copper stress responses suggesting subfunctionalization of the ancestral bifunctional protein Haa1/Cup2. *BMC Genomics* 2017, DOI: 10.1186/s12864-016-3443-2.

Palma M, Guerreiro JF, Sá-Correia I. Adaptive response and tolerance to acetic acid in *Saccharomyces cerevisiae* and *Zygosaccharomyces bailii*: a physiological genomics perspective. *Front Microbiol* 2018, DOI: 10.3389/fmicb.2018.00274.

Pan P, Hua Q. Reconstruction and in silico analysis of metabolic network for an oleaginous yeast, *Yarrowia lipolytica*. *PLoS One* 2012, DOI: 10.1371/journal.pone.0051535.

Ratledge C, Wynn JP. The biochemistry and molecular biology of lipid accumulation in oleaginous microorganisms. *Adv Appl Microbiol*. 2002. DOI: 10.1016/s0065-2164(02)51000-5.

Ravikrishnan A, Raman K. Critical assessment of genome-scale metabolic networks: the need for a unified standard. *Brief Bioinform* 2015, DOI: 10.1093/bib/bbv003.

Saier MHJ, Reddy VS, Tsu BV *et al*. The Transporter Classification Database (TCDB): recent advances. *Nucleic Acids Res* 2016;**44**:D372–9.

Schmidt FR. Recombinant expression systems in the pharmaceutical industry. *Appl Microbiol Biotechnol* 2004;**65**:363–72.

Shen XX, Opulente DA, Kominek J *et al*. Tempo and mode of genome evolution in the budding yeast subphylum. *Cell* 2018, DOI: 10.1016/j.cell.2018.10.023.

Sohn SB, Graf AB, Kim TY *et al*. Genome-scale metabolic model of methylotrophic yeast *Pichia pastoris* and its use for in silico analysis of heterologous protein production. *Biotechnol J* 2010;**5**:705–15.

Sohn SB, Kim TY, Lee JH *et al*. Genome-scale metabolic model of the fission yeast *Schizosaccharomyces pombe* and the reconciliation of in silico/in vivo mutant growth. *BMC Syst Biol* 2012;**6**:49.

Solieri L, Chand Dakal T, Croce MA *et al*. Unravelling genomic diversity of *Zygosaccharomyces rouxii* complex with a link to its life cycle. *FEMS Yeast Res* 2013, DOI: 10.1111/1567-1364.12027.

Sánchez BJ, Nielsen J. Genome scale models of yeast: towards standardized evaluation and consistent omic integration. *Integr Biol (United Kingdom)* 2015, DOI: 10.1039/c5ib00083a.

The UniProt Consortium. UniProt: the universal protein knowledgebase. *Nucleic Acids Res* 2017;**45**:D158–69.

Tiukova IA, Prigent S, Nielsen J *et al*. Genome-scale model of *Rhodotorula toruloides* metabolism. *Biotechnol Bioeng* 2019, DOI: 10.1002/bit.27162.

Tomàs-Gamisans M, Ferrer P, Albiol J. Fine-tuning the *P. pastoris* iMT1026 genome-scale metabolic model for improved prediction of growth on methanol or glycerol as sole carbon sources. *Microb Biotechnol* 2018;**11**:224–37.

Tomàs-Gamisans M, Ferrer P, Albiol J. Integration and validation of the genome-scale metabolic models of *Pichia pastoris*: a comprehensive update of protein glycosylation pathways, lipid and energy metabolism. *PLoS One* 2016, DOI: 10.1371/journal.pone.0148031.

Wang H, Marcišauskas S, Sánchez BJ *et al*. RAVEN 2.0: A versatile toolbox for metabolic network reconstruction and a case study on *Streptomyces coelicolor*. *PLoS Comput Biol* 2018;**14**:e1006541.

Wichmann G, Hawkins KM, Jackson P *et al*. Rewriting yeast central carbon metabolism for industrial isoprenoid production. *Nature* 2016;**537**:694–7.

Xu N, Liu L, Zou W *et al*. Reconstruction and analysis of the genome-scale metabolic network of *Candida glabrata*. *Mol Biosyst* 2013, DOI: 10.1039/c2mb25311a.

Ye R, Huang M, Lu H *et al*. Comprehensive reconstruction and evaluation of *Pichia pastoris* genome-scale metabolic model that accounts for 1243 ORFs. *Bioresour Bioprocess* 2017, DOI: 10.1186/s40643-017-0152-x.

Zomorrodi AR, Maranas CD. Improving the iMM904 *S. cerevisiae* metabolic model using essentiality and synthetic lethality data. *BMC Syst Biol* 2010;**4**:178.

Paper III:

Reconstruction of a catalogue of genome-scale metabolic models with enzymatic constraints using GECKO 2.0

Domenzain, I., Sánchez, B., Anton, M., Kerkhoven, E. J., Millán-Oropeza, A., Henry, C., Siewers, V., Morrisey, J. P., Sonnenschein, N. and Nielsen, J.

Nature Communications, 2023

nature COMMUNICATIONS

ARTICLE

https://doi.org/10.1038/s41467-022-31421-1 OPEN

Reconstruction of a catalogue of genome-scale metabolic models with enzymatic constraints using GECKO 2.0

Iván Domenzain[1,2], Benjamín Sánchez[3,4,9], Mihail Anton[1,5,9], Eduard J. Kerkhoven[1,2], Aarón Millán-Oropeza[6], Céline Henry[6], Verena Siewers[1,2], John P. Morrissey[7], Nikolaus Sonnenschein[3] & Jens Nielsen[1,2,8]✉

Genome-scale metabolic models (GEMs) have been widely used for quantitative exploration of the relation between genotype and phenotype. Streamlined integration of enzyme constraints and proteomics data into such models was first enabled by the GECKO toolbox, allowing the study of phenotypes constrained by protein limitations. Here, we upgrade the toolbox in order to enhance models with enzyme and proteomics constraints for any organism with a compatible GEM reconstruction. With this, enzyme-constrained models for the budding yeasts *Saccharomyces cerevisiae*, *Yarrowia lipolytica* and *Kluyveromyces marxianus* are generated to study their long-term adaptation to several stress factors by incorporation of proteomics data. Predictions reveal that upregulation and high saturation of enzymes in amino acid metabolism are common across organisms and conditions, suggesting the relevance of metabolic robustness in contrast to optimal protein utilization as a cellular objective for microbial growth under stress and nutrient-limited conditions. The functionality of GECKO is expanded with an automated framework for continuous and version-controlled update of enzyme-constrained GEMs, also producing such models for *Escherichia coli* and *Homo sapiens*. In this work, we facilitate the utilization of enzyme-constrained GEMs in basic science, metabolic engineering and synthetic biology purposes.

[1] Department of Biology and Biological Engineering, Chalmers University of Technology, SE-412 96 Gothenburg, Sweden. [2] Novo Nordisk Foundation Center for Biosustainability, Chalmers University of Technology, SE-412 96 Gothenburg, Sweden. [3] Department of Biotechnology and Biomedicine, Technical University of Denmark, 2800 Kongens Lyngby, Denmark. [4] Novo Nordisk Foundation Center for Biosustainability, Technical University of Denmark, 2800 Kongens Lyngby, Denmark. [5] Department of Biology and Biological Engineering, National Bioinformatics Infrastructure Sweden, Science for Life Laboratory, Chalmers University of Technology, Kemivägen 10, SE-412 58 Gothenburg, Sweden. [6] Plateforme d'analyse protéomique Paris Sud-Ouest (PAPPSO), INRAE, MICALIS Institute, Université Paris-Saclay, 78350 Jouy-en-Josas, France. [7] School of Microbiology, Environmental Research Institute and APC Microbiome Ireland, University College Cork, T12 K8AF Cork, Ireland. [8] BioInnovation Institute, Ole Maaløes Vej 3, 2200 Copenhagen, Denmark. [9] These authors contributed equally: Benjamín Sánchez, Mihail Anton. ✉email: nielsenj@chalmers.se

Genome-scale metabolic models (GEMs) have become an established tool for systematic analyses of metabolism for a wide variety of organisms[1–6]. Their myriads of applications span from model-driven development of efficient cell factories[3,7–9], to their utilization for understanding mechanisms underlying complex human diseases[10–12]. One of the most common simulation techniques for enabling phenotype predictions with these models is flux balance analysis (FBA), which assumes that there is balancing of fluxes around each metabolite in the metabolic network. This means that fluxes are constrained by stoichiometries of the biochemical reactions in the network, and that cells have evolved in order to operate their metabolism according to optimality principles[13,14]. Quantitative determination of biologically meaningful flux distribution profiles is a major challenge for constraint-based methods, as optimal phenotypes can be attained by alternate flux distribution profiles[15], caused by the presence of network redundancies that provide organisms with robustness to environmental and genetic perturbations. This limitation is often addressed by incorporation of experimental measurements of exchange fluxes (secretion of byproducts and uptake of substrates) as numerical flux constraints for the FBA problem. However, such measurements are not readily available for a wide variety of conditions and organisms.

In order to overcome these limitations, the concept of enzymatic limitations on metabolic reactions has been explored and incorporated by several constraint-based methods. Some of these have modeled enzyme demands of metabolic reactions by constraining metabolic networks with kinetic parameters and physiological limitations of cells, such as a crowded intracellular volume[16–18], a finite membrane surface area for expression of transporter proteins[19] and a bounded total protein mass available for metabolic enzymes[20–25]. All of these modeling frameworks have been successful at expanding the range of predictions of classical FBA, providing explanations for overflow metabolism and cellular growth on diverse environments for *Escherichia coli*[16–19,21,23,25], *Saccharomyces cerevisiae*[22,25,26], *Lactococus lactis*[27], and even human cells[20,24]. However, these modeling approaches were applied to metabolic networks of extensively studied model organisms, which are usually well represented in specialized resources for kinetic parameters such as the BRENDA[28] and SABIO RK[29] databases. Furthermore, collecting the necessary parameters for the aforementioned models was mostly done manually; therefore, no generalized model parameterization procedure was provided as an integral part of these methods.

Enzyme limitations have also been introduced into models of metabolism by other formalisms, for instance, Metabolic and gene Expression models (ME-models), implemented on reconstructions for *E. coli*[30–33], *Thermotoga maritima*[34] and *Lactococus lactis*[35]; and resource balance analysis models (RBA), on reconstructions for *E. coli*[36] and *Bacillus subtilis*[36,37]. These formalisms succeeded at merging genome-scale metabolic networks together with comprehensive representations of macromolecular expression processes, enabling detailed exploration of the constraints that govern cellular growth on diverse environments. Despite the great advances for understanding cell physiology provided by these modeling formalisms, accuracy on phenotype predictions is compromised by the large number of parameters that are required (rate constants for transcriptional, translational, protein folding and degradation processes), with most of these not being readily available in the literature. Moreover, these models encompass processes that differ radically in their temporal scales (e.g., protein synthesis vs. metabolic rates) and their mathematical representation (presence of non-linear expressions in ME-models), requiring the implementation of more elaborate techniques for numerical simulation.

GECKO, a method for enhancement of GEMs with Enzymatic Constraints using Kinetic and Omics data, was developed in 2017 and applied to the consensus GEM for *S. cerevisiae*, Yeast7[38]. This method extends the classical FBA approach by incorporating a detailed description of the enzyme demands for the metabolic reactions in a network, accounting for all types of enzyme-reaction relations, including isoenzymes, promiscuous enzymes and enzymatic complexes. Moreover, GECKO enables direct integration of proteomics abundance data, if available, as constraints for individual protein demands, represented as enzyme usage pseudo-reactions, whilst all the unmeasured enzymes in the network are constrained by a pool of remaining protein mass. Additionally, this method incorporates a hierarchical and automated procedure for retrieval of kinetic parameters from the BRENDA database, which yielded a high coverage of kinetic constraints for the *S. cerevisiae* network. The resulting enzyme-constrained model, ecYeast7, was used for successful prediction of the Crabtree effect in wild-type and mutant strains of *S. cerevisiae* and cellular growth on diverse environments and genetic backgrounds, but also provided a simple framework for prediction of protein allocation profiles and study of proteomics data in a metabolic context. Furthermore, the model formed the basis for modeling yeast growth at different temperatures[39].

Since the first implementation of the GECKO method[38], its principles of enzyme constraints have been incorporated into GEMs for *B. subtilis*[40], *E. coli*[41], *B. coagulans*[42], *Streptomyces coelicolor*[43] and even for diverse human cancer cell-lines[2], showing the applicability of the method even for non-model organisms. Despite the rapid adoption of the method by the constraint-based modeling community, there is still a need for automating the model generation and enabling identification of kinetic parameters for less studied organisms.

In this work, we updated the GECKO toolbox to its 2.0 version, expanding its use it for building enzyme-constrained models (ecModels) for more organisms. Among other improvements, we generalized its structure to facilitate its applicability to a wide variety of GEMs, and we improved its parameterization procedure to ensure high coverage of kinetic constraints, even for poorly studied organisms. Additionally, we incorporated simulation utility functions, and developed an automated pipeline for updating ecModels, named ecModels container. This container is directly connected to the original sources of version-controlled GEMs and the GECKO toolbox, offering a continuously updated catalog of diverse ecModels.

Results

Community development of GECKO. To ensure wide application and enable future development by the research community, we established the GECKO toolbox as open-source software, mostly encoded in MATLAB. It integrates modules for enhancement of GEMs with kinetic and proteomics constraints, automated retrieval of kinetic parameters from the BRENDA database (python module), as well as simulation utilities and export of ecModel files compatible with both the COBRA toolbox[44] and the COBRApy package[45]. The development of GECKO has been continuously tracked in a public repository (https://github.com/SysBioChalmers/GECKO) since 2017, providing a platform for open and collaborative development. The generation of output model files in.txt and SBML L3V1 FBC2[46] formats enabled the use of the ecYeastGEM[1] structure as a standard test to track the effects of any modifications in the toolbox algorithm through the use of the Git version control system, contributing to reproducibility of results and backwards compatibility of code.

Interaction with users of the GECKO toolbox and the ecYeastGEM model has also been facilitated through the use of the GECKO repository, allowing users to raise issues related with the programming of the toolbox or even about conceptual assumptions of the method, which has guided cumulative enhancements. Additionally, technical support for installation and utilization of the toolbox and ecYeastGEM is now provided through an open community chat room (available at: https://gitter.im/SysBioChalmers/GECKO), reinforcing transparent and continuous communication between users and developers.

New additions to the GECKO toolbox. The previous implementation of the GECKO method in GECKO 1.0 significantly improved phenotype predictions for *S. cerevisiae*'s metabolism under a wide variety of genetic and environmental perturbations[38]. However, its development underscored some issues, in particular that quantitative prediction of the critical dilution rate and exchange fluxes at fermentative conditions are highly sensitive to the distribution of incorporated kinetic parameters. Although *S. cerevisiae* is one of the most studied eukaryote organisms, not all reactions included in its model have been kinetically characterized. Therefore, a large number of k_{cat} numbers measured for other organisms (48.35%), or even non-specific to their reaction mechanism (56.03% of k_{cat} values found by introduction of wildcards into E.C. numbers) were needed to be incorporated, in order to fill the gaps in the available data for the reconstruction of the first *S. cerevisiae* ecModel, ecYeast7. Moreover, detailed manual curation of k_{cat} numbers was needed for several key enzymes in order to achieve biologically meaningful predictions.

As the BRENDA database[47] is the main source of kinetic parameters for GECKO, all of the available k_{cat} and specific activity entries for non-mutant enzymes were retrieved. In total, 38,280 entries for 4130 unique E.C. numbers were obtained and classified according to biochemical mechanisms, phylogeny of host organisms and metabolic context (Brenda kinetic data analysis section in the Supplementary Information File 1), in order to assess significant differences in distributions of kinetic parameters. This analysis showed that not all organisms have been equally studied. While entries for *H. sapiens, E. coli, R. norvegicus*, and *S. cerevisiae* account for 24.02% of the total, very few kinetic parameters are available for most of the thousands of organisms present in the database, showing a median of 2 entries per organism (Fig. 1a). The analysis also showed that kinetic activity can differ drastically, spanning several orders of magnitude even for families of enzymes with closely related biochemical mechanisms (Fig. 1b). Finally, it was also observed that k_{cat} distributions for enzymes in the central carbon and energy metabolism differ significantly from those in other metabolic contexts across phylogenetic groups of host organisms (life kingdoms, according to the KEGG phylogenetic tree[48]), even without filtering the dataset for entries reported exclusively for natural substrates, as previously done by other studies[49] (Fig. 1c).

In the new version of the GECKO toolbox (GECKO 2.0), a modified set of hierarchical k_{cat} matching criteria was implemented to address how k_{cat} numbers depend on biochemical mechanisms, metabolic context and phylogeny of host organisms. The modified parameterization procedure enables the incorporation of kinetic parameters that have been reported as *specific activities* in BRENDA when no k_{cat} is found for a given query (as the specific activity of an enzyme is defined as its k_{cat} over its molecular weight), adding 8,118 new entries to the catalog of kinetic parameters in the toolbox. A phylogenetic distance-based criterion, based on the phylogenetic tree available in the KEGG database[48], was introduced for cases in which no organism-specific entries are available for a given query in the kinetic parameters dataset. Specifically, where GECKO 1.0 chooses k_{cat} available in BRENDA regardless of organism, GECKO 2.0 chooses the values available in BRENDA for the phylogenetically closest organism by iteratively introducing a wildcard into the E.C. number, as exemplified in the Brenda kinetic data analysis section in the Supplementary Information File 1 "EC3.x.x.x", and estimating the phylogenetic distance. The new k_{cat} matching algorithm, including the estimation of the phylogenetic distance, and its comparison with the predecessor are shown in the supplementary methods section in Supplementary File 1.

In order to assess the impact of the modified k_{cat} assignment algorithm on an ecModel, ecYeast7 was reconstructed using both the first and GECKO 2.0. A classification of the matched k_{cat} values according to the new matching algorithm is provided in Fig. 1d, showing the amount of values chosen from the phylogenetically closest organisms. The incorporation of specific activity values in the parameter catalog increased the number of kinetic parameters matched to complete E.C. numbers (no added wildcards) from 1432 to 2696 (Fig. 1e). Moreover, the implementation of the phylogenetic distance-based criterion yielded a distribution of kinetic parameters that showed no significant differences when compared to the values reported in BRENDA for all fungi species, in contrast to the kinetic profile matched by the previous algorithm (P-values 2.1×10^{-11} and 3.9×10^{-8}, when compared to the BRENDA fungi and *S. cerevisiae* distributions, respectively, under a two-tailed Kolmogorov–Smirnov test) (Fig. 1f). The quality of phenotype predictions for the ecYeast7 model enhanced by GECKO 2.0 was evaluated by simulation of batch growth in 19 different environments, with an average relative error of 23.97% when compared to experimental data (Fig. 1g); in contrast, its GECKO 1.0 counterpart yielded an average relative error of 32.07%.

The introduction of manually curated k_{cat} numbers in a metabolic network has been proven to increase the quality of phenotype predictions for *S. cerevisiae*[22,25,38]; nevertheless, this is an intensive and time-consuming procedure that is hard to ensure for a large number of models subject to continuous modifications. In order to ensure applicability of the GECKO method to any standard GEM, a unified procedure for curation of kinetic parameters was developed based on parameter sensitivity analysis. For automatically generated ecModels that are not able to reach the provided experimental value for maximum batch growth rate, an automatic module performs a series of steps in which the top enzymatic limitation on growth rate is identified through the quantification of enzyme control coefficients. For such enzymes, the E.C. number is obtained and then its correspondent k_{cat} value is substituted by the highest one available in BRENDA for the given enzyme class. This procedure iterates until the specific growth rate predicted by the model reaches the provided experimental value.

Finally, as the first version of the toolbox relied on the structure and nomenclature of the model Yeast7, its applicability to other reconstructions was not possible in a straightforward way. In order to provide compatibility with any other GEM, based on COBRA[44] or RAVEN[50] formats, all of the organism-specific parameters required by the method (experimental growth rate, total protein content, organism name, names and identifiers for some key reactions, etc.) can be provided in a single MATLAB initialization script, minimizing the modifications needed for the generation of a new ecModel.

ecModels container is an automatically updated repository. Several GEMs that have been published are still subject to continuous development and maintenance[1–3,5,6], this renders GEMs

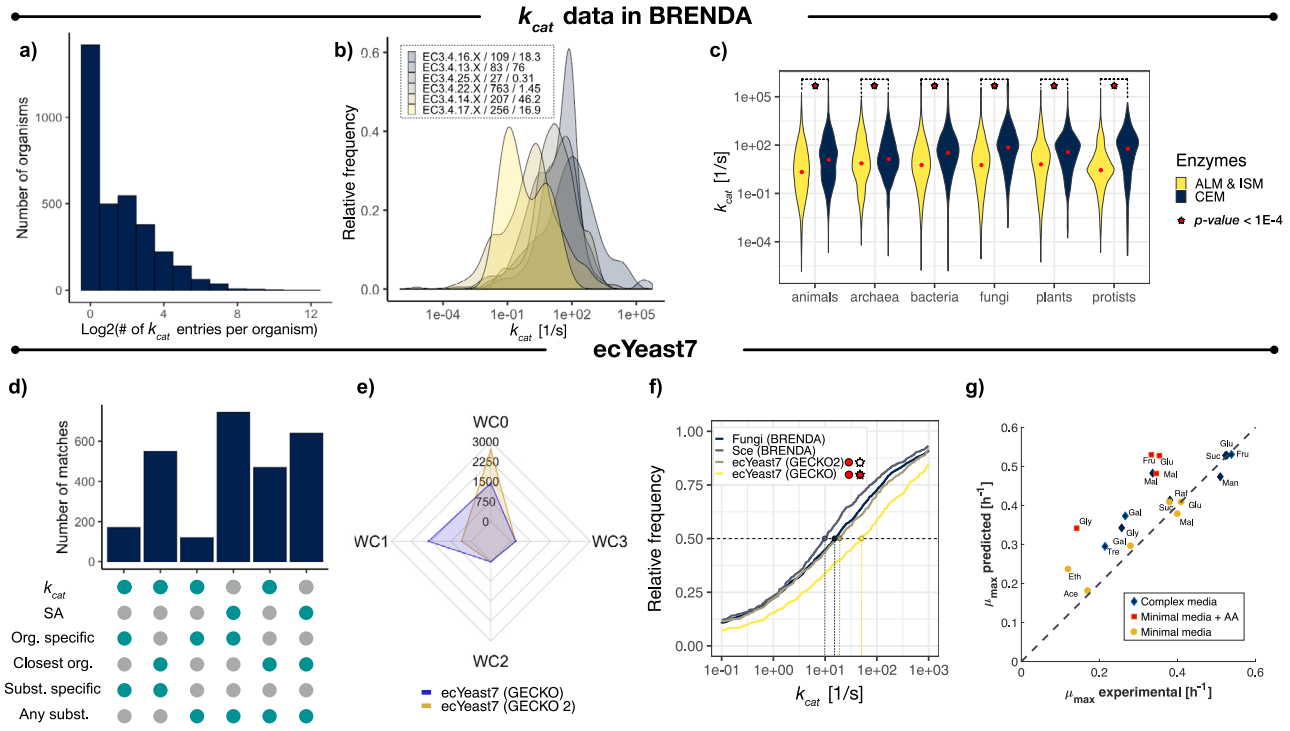


Fig. 1 k_{cat} distributions in BRENDA and ecYeast7. a Number of k_{cat} entries in BRENDA per organism. **b** k_{cat} distributions for closely related enzyme families. Sample size and median values (in s^{-1}) are shown after each family identifier. **c** k_{cat} distributions for enzymes in BRENDA by metabolic context and life kingdoms. Median values are indicated by red dots in each distribution, statistical significance (under a one-sided Kolmogorov–Smirnov test) is indicated by red stars for each pair of distributions for a given kingdom. CEM—central carbon and energy metabolism; ALM—Amino acid and lipid metabolism; ISM—intermediate and secondary metabolism. Computed *P*-values are 2.8×10^{-27} for animals; 3.85×10^{-5} for archaea; 1.62×10^{-92} for bacteria; 1.024×10^{-30} for fungi; 2.36×10^{-16} for plants and 4.75×10^{-21} for protists. **d** Number of k_{cat} matches in ecYeast7 per assignment category (GECKO 2.0). **e** Comparison of the number of k_{cat} matches for E.C. numbers with 0, 1, 2, and 3 introduced wildcards by GECKO 2.0 and GECKO k_{cat} matching algorithms. **f** Cumulative k_{cat} distributions for: all *S. cerevisiae* entries in BRENDA, all entries for fungi in BRENDA, ecYeast7 enhanced by GECKO and ecYeast7 enhanced by GECKO 2.0. Colored points and vertical dashed lines indicate the median value for each distribution. Statistical significance under a two-sided Kolmogorov–Smirnov test of the matched k_{cat} distributions when compared to all entries for *S. cerevisiae* and fungi, is shown with red circles and stars, respectively. *P*-values below 1×10^{-2} are indicated with red. Computed *P*-values are 0.538 for the comparison between GECKO2 vs. all fungi, 2.7×10^{-3} for GECKO2 vs. *S. cerevisiae*, 3.9×10^{-8} for GECKO vs. all fungi and, 2.1×10^{-11} for GECKO vs. the *S. cerevisiae* entries. **g** Prediction of batch maximum growth rates on diverse media with ecYeast7 enhanced by GECKO 2.0. Glu—glucose, Fru—fructose, Suc—sucrose, Raf— raffinose, Mal—maltose, Gal—galactose, Tre—trehalose, Gly—glycerol, Ace—acetate, Eth —ethanol. Source data are provided in Source Data: Data Source file 1.

to be dynamic structures that can change rapidly. In order to integrate such continuous updates into the enzyme-constrained version of a model in an organized way, an automated pipeline named *ecModels container* was developed.

The ecModels container is a continuous integration implementation whose main functionality is to provide a catalog of ecModels for several relevant organisms that are automatically updated every time a modification is detected either in the original GEM source repository or in the GECKO toolbox, i.e., new releases in their respective repositories. The pipeline generates ecModels in different formats, including the standard SBML and MATLAB files, and stores them in a container repository (https://github.com/SysBioChalmers/ecModels) in a version-controlled way, requiring minimal human interaction and maintenance. The GECKO toolbox ensures the creation of functional and calibrated ecModels that are compatible with the provided experimental data (maximum batch growth rate, total protein content of cells, and exchange fluxes at different dilution rates as an optional input). This whole computational pipeline is illustrated in Fig. 2. Further description of the ecModels container pipeline functioning is included in the "Methods" section.

A catalog of new ecModels. Following the aforementioned additions to the GECKO toolbox, that have allowed its generalization, we used the toolbox for the reconstruction of four new ecModels from previously existing high-quality metabolic network reconstructions: *i*Yali4, for the oleaginous yeast *Yarrowia lipolytica*[5]; *i*SM996, for the thermotolerant yeast *Kluyveromyces marxianus*[6]; *i*ML1515, for the widely studied bacterium *E. coli*[4]; and Human1, being the latest and largest network reconstruction available for studying *H. sapiens* metabolism[2]. For the microbial models, all model parameters were calibrated according to the provided experimental data, generated by independent studies[4,51–53], yielding functional ecModels ready for simulations. Size metrics for these models can be seen in Table 1.

These ecModels, together with ecYeastGEM, are hosted in the ecModels container repository for their continuous and automated update every time that a version change is detected either in the original model source or in the GECKO repository. In the case of microbial species, two different model structures are provided: *ecModel*, which has unbounded individual enzyme usage reactions ready for incorporation of proteomics data; and *ecModel_batch* in which all enzyme usage reactions are connected to a shared protein pool. This pool is then constrained by experimental values of total protein content, and calibrated for batch simulations using experimental measurements of maximum batch growth rates on minimal glucose media, thus providing a functional ecModel structure ready for simulations.

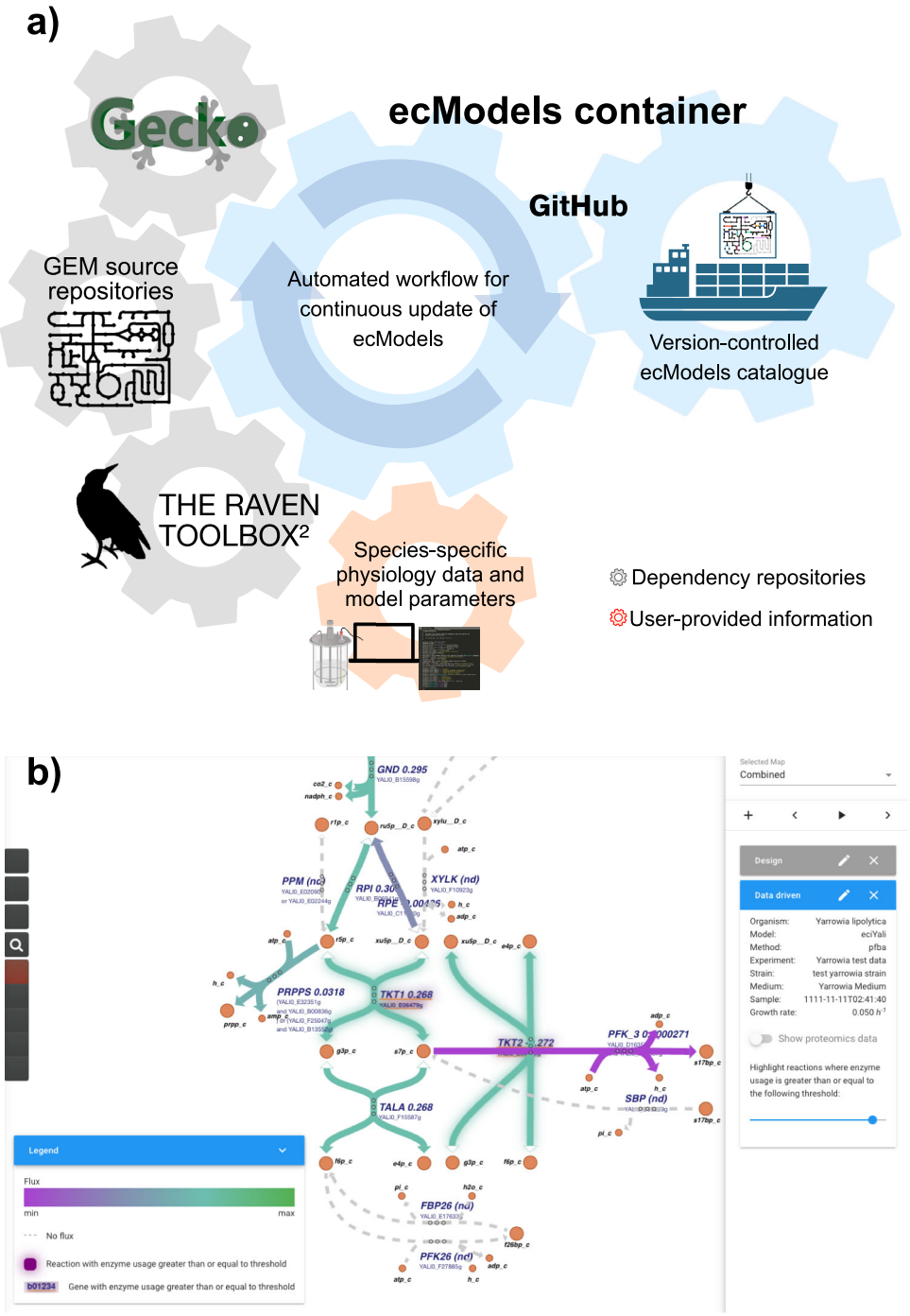


Fig. 2 Extending utilization of ecModels. a ecModels container: Integrated pipeline for continuous and automated update of ecModels. **b** Implementation of GECKO simulations in the Caffeine platform (https://caffeine.dd-decaf.eu/) for visualization of enzyme usage. The color of the arrows corresponds to the value of the corresponding fluxes. Genes or reactions connected to enzymes with a usage above 90% are highlighted with a glow around the corresponding text or arrow, respectively. The chosen usage threshold to highlight can be tuned with the slider on the right.

For ecHumanGEM just the unbounded ecModel files are provided, as this is a general network of human metabolism, containing all reactions from any kind of human tissue or cell type for which evidence is available, and therefore not suitable for numerical simulation. As *H. sapiens* is the most represented organism in the BRENDA database, accounting for 11% of the total number of available k_{cat} values (Brenda kinetic data analysis section in the Supplementary Information File 1), kinetic parameters from other organisms were not taken into account for its enhancement with enzyme constraints. ecHuman1 provides the research community with an extensive knowledge base that represents a complete and direct link between genes, proteins, kinetic parameters, reactions and metabolites for human cells in a single model structure, subject to automated continuous update by the *ecModels container* pipeline.

Visualization of GECKO simulations in the Caffeine platform. We implemented simulations with ecModels in Caffeine, an open-source software platform for cell factory design. Caffeine, publicly available at http://caffeine.dd-decaf.eu, allows user-

Table 1 Size metrics summary for the ecModels catalog.

Original GEMs					
Organism	*S. cerevisiae*	*Y. lipolytica*	*K. marxianus*	*E. coli*	*H. sapiens*
Model ID	yeastGEM_8.3.3	*i*Yali4	*i*SM996	*i*ML1515	Human1
Reactions	3963	1924	1913	2711	13101
Metabolites	2691	1671	1531	1877	8400
Genes	1139	847	996	1516	3628
Enzyme-constrained GEMs					
Model ID	ecYeastGEM	eci*Yali*	eci*SM996*	eci*ML1515*	ecHumanGEM
Reactions	8028	3881	5334	6084	46259
Metabolites	4153	1880	2064	2334	12191
Enzymes	965	647	716	1259	3224
Enzyme coverage	84.72%	76.39%	71.89%	83.05%	88.86%
Reactions w/k_{cat}	3771	1586	2891	2562	27014
Reactions w/ Isoenzymes	504	205	532	456	3791
Promiscuous Enzymes	572	324	469	673	2184
Enzyme complexes	252	75	27	383	756

friendly simulation and visualization of flux predictions made by genome-scale metabolic models. Several standard modeling methods are already included in the platform, such as ^{13}C fluxomics data integration, and simulation of gene deletion and/or overexpression, to interactively explore strain engineering strategies. In order to allow for GECKO simulations, we added a new feature to the platform for uploading enzyme-constrained models and absolute proteomics data. Additionally, we added a simulation algorithm that recognizes said models, and overlays the selected proteomics data on them, leaving out data that makes the model unable to grow at a pre-specified growth rate. After these inclusions to the platform, enzyme usage can now be computed on the fly and visualized on metabolic maps (Fig. 2b), to identify potential metabolic bottlenecks in a given condition. The original proteomics data can be visualized as well, to identify if the specific bottleneck is due to a lack of enzyme availability, or instead due to an inefficient kinetic property. This will suggest different metabolic engineering strategies to the user: if the problem lies in the intracellular enzyme levels, the user can interpret this as a recommendation for overexpressing the corresponding gene, whereas if the problem lies in the enzyme efficiency, the user could assess introducing a heterologous enzyme as an alternative.

GECKO simulation utilities. As ecModels are defined in an irreversible format and incorporate additional elements such as enzymes (as new pseudo-metabolites) and their usages (represented as pseudo-reactions), they might sometimes not be directly compatible with all of the functionalities offered by currently available constraint-based simulation software[44,45,50,54,55]. We therefore added several new features to the GECKO toolbox that allow the exploration and exploitation of ecModels. These include utilities for: (1) basic simulation and analysis purposes, (2) accessible retrieval of kinetic parameters, (3) automated generation of condition-dependent ecModels with proteomic abundance constraints, (4) comparative flux variability analysis between a GEM and its ecModel counterpart, and (5) prediction of metabolic engineering targets for enhanced production with an implementation of the FSEOF method[56] for ecModels. Detailed information about the inputs and outputs for each utility can be found on their respective documentation, available at: https://github.com/SysBioChalmers/GECKO/tree/master/geckomat/utilities. All of these utilities were developed in MATLAB due to their dependency on some RAVEN toolbox functions[50].

Predicting microbial proteome allocation in multiple environments. In order to test the quality of the phenotype predictions of an ecModel automatically generated by the *ecModels container* pipeline, batch growth under 11 different carbon sources was simulated with eci*ML1515* for *E. coli*. Figure 3a shows that, for all carbon sources, growth rates were predicted at the same order of magnitude as their corresponding experimental measurements, with the most accurate predictions obtained for growth on D-glucose, mannose and D-glucosamine. Furthermore, batch growth rate and protein allocation predictions, using no exchange flux constraints, were compared between eci*ML1515* and the *i*JL1678 ME-model[32], the latter accounting for both metabolism and macromolecular expression processes. The sum squared error (SSE) for batch growth rate predictions across the 11 carbon sources using eci*ML1515* was 0.27, a drastic improvement when compared to the 1.21 SSE of *i*JL1678 ME-model predictions[32]. Figure 3b shows the predicted total proteome needed by cells to sustain the provided experimental growth rates for the same 11 environments. Notably eci*ML1515* predicts values that lie within the range of predictions of the *i*JL1678 ME-model (from the optimal to the generalist case) for 10 out of the 11 carbon sources (see "Methods" for simulation details). This shows that the new version of the GECKO toolbox ensures the generation of functional ecModels that can be readily used for simulation of metabolism, due to its systematic parameter flexibilization step, which reduces the need of extensive manual curation for new ecModels. Furthermore, *i*ML1515 is a model available as a static file at the BiGG models repository[57]; therefore, its integration to the ecModels container for continuous update demonstrates the flexibility of our pipeline, regarding compatibility with original GEM sources, which can be provided as a link to their *git*-based repositories or even as static URLs.

Proteomics constraints refine phenotype predictions for multiple organisms and conditions. The previously mentioned module for integration of proteomics data generates a condition-dependent ecModel with proteomics constraints for each condition/replicate in a provided dataset of absolute protein abundances [mmol/gDw]. Even though absolute quantification of proteins is becoming more accessible and integrated into systems biology studies[58–62], a major caveat of using proteomics data as constraints for quantitative models is their intrinsic high biological and technical variability[63], therefore some of the incorporated data constraints need to be loosened in order to obtain functional ecModels. When needed, additional condition-

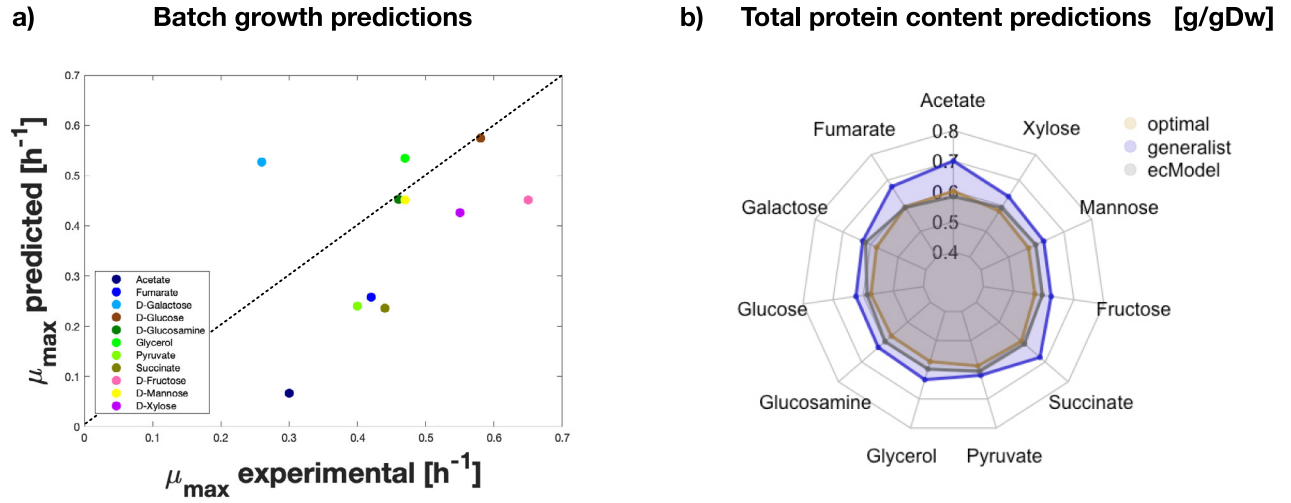


Fig. 3 Comparison of predictive capabilities between eci*ML1515* and ME-*iJL1678* for *E. coli*. **a** Maximum batch growth rate predictions on minimal media with diverse carbon sources, with an average relative error for eci*ML1515* of 34,43%, and an R^2 of 0.196. The sum of squared errors when compared to experimental values are 0.2785 for eci*ML1515* and 1.21 for ME-*iJL1678*. **b** Prediction of total protein content in the cell by eci*ML1515* and ME-*iJL1678* using the optimal and generalist approaches. Source data are provided in Source Data: Data Source file 1.

dependent exchange fluxes of byproducts can also be used as constraints in order to limit the feasible solution space. A detailed description of the proteomics integration algorithm implemented in GECKO is given in the supplementary methods section in the Supplementary Information File 1.

The new proteomics integration module was tested on the three ecModels for budding yeasts available in ecModels container (ecYeastGEM, eci*Yali*, eci*SM996*). We measured absolute protein abundances for *S. cerevisiae, Y. lipolytica* and *K. marxianus*, grown in chemostats at 0.1 h^{-1} dilution rate and subject to several experimental conditions (high temperature, low pH and osmotic stress with KCl)[64], and incorporated these data into the ecModels as upper bounds for individual enzyme usage pseudo-reactions. Then, exchange fluxes for CO_2 and oxygen corresponding to the same chemostat experiments were used as a comparison basis to evaluate quality of phenotype predictions. For each organism- condition pair, 3 models were generated and compared in terms of predictions: a pure stoichiometric metabolic model, an enzyme-constrained model with a limited shared protein pool, and an enzyme-constrained model with proteomics constraints. It was found that the addition of the enzyme pool constraint enables major reduction of the relative error in prediction of gaseous exchange fluxes in some of the studied conditions. Additionally, the incorporation of individual protein abundance constraints improves even further the predictive accuracy of gaseous exchanges, for 5 out of the 11 evaluated cases (Fig. 4a–c). Although only a trend and not a significant improvement, it would be of interest, in the future, to run further analyses that include more proteomics datasets.

The impact of incorporating enzyme and proteomics constraints on intracellular flux predictions was further assessed by mapping all condition-dependent flux distributions from the tested ecModels to their corresponding reactions in the original GEMs. In general, metabolic flux distributions showed high similarity when comparing ecModel to GEM predictions (Supplementary Fig. 1), as 70–90% of the active reaction fluxes were predicted within the interval of $0.5 < \text{fold-change} < 2$ $\left(\text{FC} = \frac{v_i^{\text{ecModel}}}{v_i^{\text{GEM}}}\right)$ across all conditions (Supplementary Fig. 2A–C, Source Data: Data Source File 2). In addition, principal component analysis on absolute enzyme usage profiles predicted by ecModels revealed that, at low dilution rates, predictions of enzyme demands are mostly defined by the selected set of imposed constraints (shared protein pool vs. proteomics constraints) rather than by environmental condition, i.e., exchange fluxes (Supplementary Fig. 2D–F). However, more straightfroward comparison of the models' predictions, by pairwise comparison of predicted absolute enzyme usage profiles, showed that 60–80% of the predicted enzyme usages lie within a range of $0.5 < \text{fold-change} < 2$, when comparing ecModels predictions with and without proteomics constraints, across organisms and conditions (Fig. 4d, Supplementary Fig. 2G–I, and Data Source File 2). It was observed that the incorporation of proteomics constraints induces a drastic differential use for a considerable amount of enzymes, as 12–21% of enzyme usages were predicted as either enabled or disabled by these constraints across all the simulated conditions, showing slight enrichment for enabled alternative isoenzymes for already active reactions (Data Source File 2). This suggests that upper bounds on enzyme usages induce differentiated utilization of isoenzymes, reflecting well why isoenzymes have been maintained throughout evolution.

The explicit inclusion of enzymes into GEMs by the GECKO method enables prediction of enzyme demands at the protein, reaction and pathway levels. Total protein burden values predicted by ecModels for several relevant metabolic superpathways (central carbon and energy metabolism, amino acid metabolism, lipid and fatty acid metabolism, cofactor and vitamin metabolism and nucleotide metabolism, according to the KEGG metabolic subsystems[48]), showed that central carbon and energy metabolism is the most affected sector in the ecYeastGEM network by integration of proteomics constraints, as protein burden predictions were higher, at least by 20%, for 3 out of the 4 simulated conditions when compared with predictions of the ecYeastGEM without proteomics data (Fig. 4e).

Relative enzyme usages, estimated as predicted absolute enzyme usage over enzyme abundance for all of the measured enzymes in an ecModel $\left(\frac{e_i}{[E_i]}\right)$, can be understood as the saturation level of enzymes in a given condition. In order to analyze the metabolic mechanisms underlying long-term adaptation to stress in budding yeasts, relative enzyme usage profiles

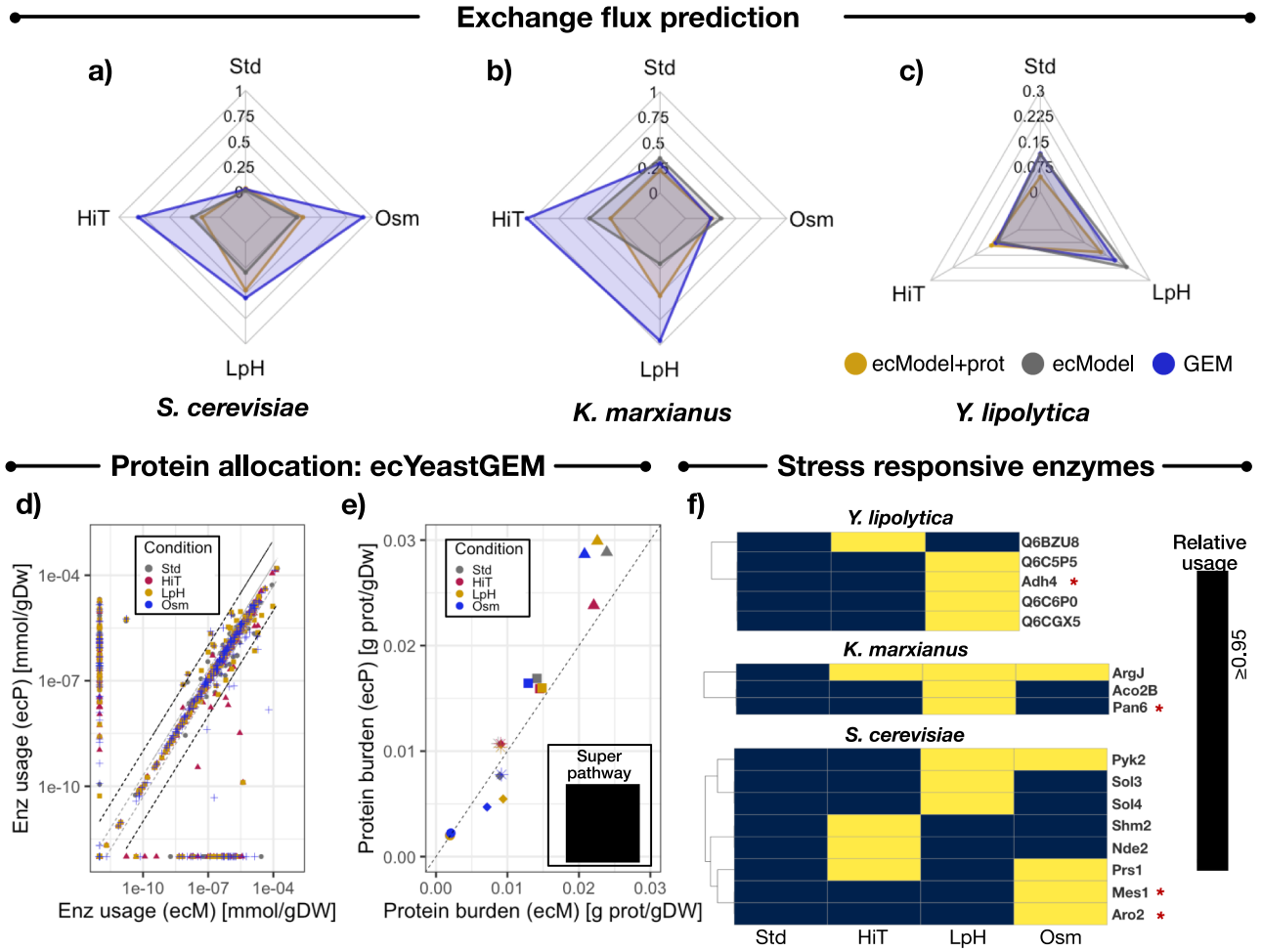


Fig. 4 Evaluation of proteomics-constrained ecModels. Comparison of median relative error in prediction of exchange fluxes for O_2 and CO_2 by GEMs, ecModels and proteomics-constrained ecModels across diverse conditions (chemostat cultures at 0.1 h^{-1} dilution rate) for **a** *S. cerevisiae*, **b** *K. marxianus*, **c** *Y. lipolytica*. **d** Comparison of absolute enzyme usage profiles [mmol/gDw] predicted by ecYeastGEM (ecM) and ecYeastGEM with proteomics constraints (ecP) for several experimental conditions. The region between the two dashed gray lines indicates enzyme usages predicted in the interval $0.5 \leq E_i^{ecP}/E_i^{ecM} \leq 2$, the region between the two dashed black lines indicates enzyme usages predicted in the interval $0.1 \leq E_i^{ecP}/E_i^{ecM} \leq 10$, when comparing the two ecModels. **e** Protein burden for different superpathways predicted by ecYeastGEM (ecM) and ecYeastGEM with proteomics constraints (ecP). **f** Highly saturated enzymes at different stress conditions for *S. cerevisiae, K. marxianus*, and *Y. lipolytica* predicted by their corresponding ecModels constrained with proteomics data. Yellow cells indicate condition-responsive enzymes (relative usage ≥ 0.95). Red asterisks indicate enzymes conserved as single copy orthologs across the three yeast species. Std—Reference condition, HiT—high-temperature condition, LpH—Low pH condition, Osm—Osmotic stress condition, AA—amino acid metabolism, NUC—nucleotide metabolism, CEM—central carbon and energy metabolism, CofVit—cofactor and vitamin metabolism, Lip—lipid and fatty acid metabolism. Source data are provided in Source Data: Data Source File 2.

were computed from all the previous simulations of ecModels with proteomics constraints. Enzymes that display fold-changes higher than 1 for both absolute abundance and their saturation level, when comparing predicted usage profiles between stress and reference conditions, suggest regulatory mechanisms on individual proteins that contribute to cell growth on the anlyzed stress condition. Figure 4f shows all of the enzymes that were identified as responsive to environmental stress in this study, displaying enrichment for enzymes involved in biosynthesis of diverse amino acids and folate metabolism.

A further mapping of all enzymes in these ecModels to a list of 2,959 single copy protein-coding gene orthologs across the three yeast species[64] found 310 core proteins across these ecModels. Principal component analysis revealed that variance on absolute enzyme usages and abundance profiles for these core proteins is mostly explained by differences in the metabolic networks of the different species rather than by environmental conditions (Supplementary Fig. 3B, C), reinforcing previous results suggesting that, despite being phylogenetically related, their long-term stress responses at the molecular level have evolved independently after their divergence in evolutionary history[64].

Exploring the solution space reduction. A major limitation in the use of GEMs is the high variability of flux distributions for a given cellular objective when implementing flux balance analysis, as this requires solving largely underdetermined linear systems through optimization algorithms[15,65]. This limitation has usually been overcome with incorporation of measured exchange fluxes as constraints. However, these data are typically sparse in the literature. Previous studies explored the drastic reduction in flux variability ranges of ecModels for *S. cerevisiae* and 11 human cell-lines when compared to their original GEMs due to the addition of enzyme constraints[1,2,38]. However, the irreversible format of ecModels (forward and backwards reactions are split in order to account for enzyme demands of both directions) hinders their compatibility with the flux variability analysis (FVA) functions

already available in COBRA[44] and RAVEN[50] toolboxes. As a solution to this, an FVA module was integrated to the utilities repertoire in GECKO, whose applicability has been previously tested on studies with ecModels for *S. cerevisiae*[1] and human cell lines[2]. This module contains the necessary functions to perform FVA on any set of reactions of an ecModel, enabling also a direct comparison of flux variability ranges between an ecModel and its GEM counterpart in a consistent way (supplementary methods section in the Supplementary Information File 1).

The FVA utility was applied on three different ecModels of microbial metabolism and their correspondent GEMs (*i*ML1515, *i*Yali4, and *i*SM996). In all cases the FVA comparisons were carried out for both chemostat and batch growth conditions in order to span different degrees of constraining of the metabolic networks (0.1 h^{-1} dilution rate and minimal glucose uptake rate fixed for chemostat conditions; biomass production fixed to experimental measurements of μ_{max} and unconstrained uptake of minimal media components, for batch conditions). Cumulative distributions for flux variability ranges for all explored ecModels and GEMs are shown in Fig. 5, in which it can be seen that median flux variability ranges are much reduced for all ecModels and conditions, especially at high growth rates where enzyme constraints reduce the variability range 5–6 orders of magnitude when compared to pure GEMs. The cumulative distributions also show a major reduction in the amount of totally variable fluxes (reactions that can carry any flux between −1000 to 1000 mmol/gDwh), which are an indicator of undesirable futile cycles present in the network due to lack of thermodynamic and enzyme cost information[66–68]. For high growth rates, the amount of totally variable fluxes accounts for 3–12% of the active reactions in the analyzed GEMs, in contrast to their corresponding ecModels in which such extreme variability ranges are completely absent.

Further analysis of the FVA results revealed that a reduction of at least 95% of the variability range was achieved for more than 90% of all fluxes of active reactions at high growth rates in all ecModel. Interestingly, the aforementioned flux variability metrics were overall improved even for the chemostat conditions, despite a higher degree of constraining (fixed low growth rate and optimal uptake rate), which restrains these models to an energy efficient respiratory mode (Data Source File 3).

Discussion

Here, we demonstrated how enzyme-constrained models for diverse species significantly improve simulation performance compared to traditional GEMs. Furthermore, to enable the community to easily adapt this modeling approach, we upgraded the GECKO toolbox for enhancement of genome-scale models with enzyme and omics constraints to its version 2.0. Major improvements on the k_{cat} matching algorithm were incorporated into the toolbox, based on phylogenetic distance between the modeled organism and the host organisms for data queries, and an automated curation of k_{cat} numbers for over-constrained models were incorporated into the toolbox. Major refactoring of the GECKO toolbox enabled a generalization of the method, allowing the creation of high-quality ecModels for any provided functional GEM with minimal need for case-specific introduction of new code. Additionally, several utility functions were integrated into the toolbox in order to enable basic simulation purposes, accessible retrieval of enzyme parameters, integration of proteomics data as constraints, flux variability analysis and prediction of gene targets for enhanced production of metabolites. Overall, it was shown that these enhancements to the GECKO toolbox improve the incorporation of kinetic parameters into a metabolic model, yielding ecModels with biologically meaningful kinetic profiles without compromising accuracy on phenotype predictions.

Two major limitations of the first version of the GECKO toolbox were its specific customization to the *S. cerevisiae* model, Yeast7, and the need of extensive manual curation for generating an ecModel suited for FBA simulations; thus, its applicability to

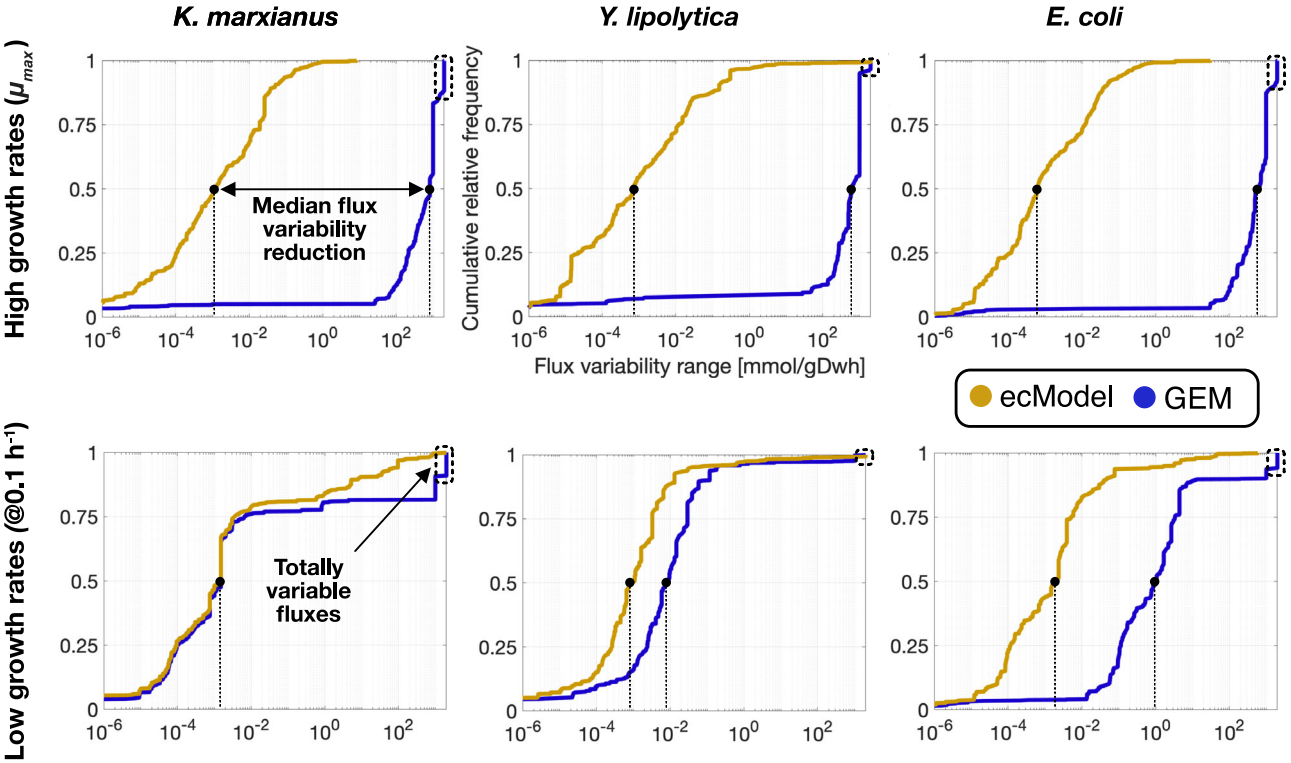


Fig. 5 Cumulative distributions of flux variability ranges for *i*SM996, *i*Yali4 and *i*ML1515 compared to their respective enzyme-constrained versions at low and high growth rates. Source data are provided in the Source Data: Data Source File 3.

other GEMs was not a straightforward procedure. To overcome these limitations, we generalized the code with the aim of making GECKO a model-agnostic tool. The development of a procedure for automatic curation of kinetic parameters enabled the generation of functional ecModels with minimal requirements for experimental data. Recently, ecModels for 11 human cancer cell-lines were generated with this automated procedure, using Human1 as a model input and RNAseq datasets together with the tINIT algorithm[10] to generate cell-line specific networks[2]. These ecModels were used for the prediction of cellular growth and metabolite exchange rates at different levels of added constraints, resulting in remarkable improvements in accuracy when compared with predictions of their original GEMs. This highlights one of the main advantages of ecModels: their capability of yielding biologically meaningful phenotype predictions without an excessive dependency on exchange fluxes as constraints.

In order to further showcase the functionality of the GECKO toolbox 2.0, a family of new high-quality ecModels were generated for *E. coli, Y. lipolytica, K. marxianus* and *H. sapiens*, based on the original GEMs *i*ML1515, *i*Yali4, *i*SM996 and Human1, respectively. Furthermore, we generated a self-hosted pipeline for continuous and automated generation and update of ecModels, *ecModels container*, so that each of the currently available ecModels (ecYeastGEM, ec*i*ML1515, ec*i*Yali, ec*i*SM996, and ecHuman1) are integrated to it, providing a version-controlled and continuously updated repository for high-quality ecModels. Moreover, the implemented automation facilitates the application of the GECKO method to other organisms for which sufficient data is available.

Absolute proteomics measurements for the budding yeasts *S. cerevisiae*, *K. marxianus* and *Y. lipolytica* grown under multiple environmental conditions, were incorporated as constraints into their ecModels by using the proteomics integration module added to the GECKO toolbox. Analysis of metabolic flux distributions revealed that net reaction fluxes predicted by GEMs are not significantly affected by the incorporation of kinetic and proteomics constraints, however, the explicit integration of enzymes into ecModels extends the range of predictions of classical FBA and enables computation of enzyme demands at the reaction and pathway levels. It was found that incorporation of proteomics constraints does not affect enzyme demand predictions significantly for most of the active enzymes at low dilution rates across the simulated conditions. However, we observed that a diversified utilization of isoenzymes, enforced by proteomics constraints, increases the predicted total protein mass allocated to central carbon and energy metabolism, in comparison to optimal enzyme allocation profiles. This result suggests the relevance of metabolic robustness in contrast to optimal protein utilization for microbial growth under environmental stress and nutrient-limited conditions.

Incorporation of proteomics data allows the use of ecModels as scaffolds for systems-level studies of metabolism, providing a tool for uncovering metabolic readjustments induced by genetic and environmental perturbations, which might be difficult to elucidate by purely data-driven approaches, specially at conditions of relatively low changes at the transcript[69] and protein levels[64]. For all studied stress conditions in this study, we identified upregulated proteins (increased abundance) that are needed to operate at high saturation levels in stress conditions, while showing low usage at reference conditions, creating lists of potential gene amplification targets for enhancing stress tolerance in three industrially relevant yeast species (Source Data: Data Source File 2). Upregulation and high saturation of enzymes in amino acid and folate metabolism were found to be common across the studied organisms and stress conditions (Supplementary Fig. 3D and Source Data: Data Source File 2). These results suggest that yeast cells display enzyme expression profiles that provide them with metabolic robustness for microbial growth under stress and nutrient-limited conditions, in contrast to an optimal protein allocation strategy that prioritizes expression of the most efficient and non-redundant enzymes.

Our results on drastic reduction of median flux variability ranges and the number of totally unbounded fluxes for ec*i*Yali, ec*i*SM996, and ec*i*ML1515, together with previous studies[1,2,38], suggest that a major reduction of the solution space of metabolic models to a more biologically meaningful subspace is a general property of ecModels. However, flux variability is an intrinsic characteristic of metabolism; therefore, metabolic models with highly constrained solution spaces may exclude some biological capabilities of organisms, which are not compatible with the set of constraints used for the analysis (exchange fluxes, growth rates and even profiles of kinetic parameters, considered as condition-independent in ecModels).

Here, the predictive capabilities of ec*i*ML1515 and *i*JL1678 ME-model (both for *E. coli*) for cellular growth and global protein demands on diverse environments were compared. The major improvement in predicted maximum growth rates, together with a comparable performance on quantification of protein demands, shown by ec*i*ML1515 suggest that, despite its mathematical and conceptual simplicity, the GECKO formalism is a suitable framework for quantitative probing of metabolic capabilities, compatible with the widely used FBA method and without the need of excessive complexity or computational power. Nevertheless, ME-models provide a much wider range of predictions that explore additional processes in cell physiology with great detail. Direct comparison between the predictions of these modeling formalisms, suggest that ME-models performance can be improved by incorporation of either curated or systematically retrieved kinetic parameters that are suitable for the modeled organisms.

Simpler modeling frameworks that account for protein or enzyme constraints in metabolism, such as flux balance analysis with molecular crowding (FBAwMC)[16,17], metabolic modeling with enzyme kinetics (MOMENT)[23], and constrained allocation flux balance analysis (CAFBA)[21], have also been developed and used to explore microbial cellular growth[16,17,21] and overflow metabolism[16,23]. These methods have overcome the lack of reported parameters for some specific reactions either by incorporation of proteomics measurements and prior flux distributions[23], manual curation and sampling procedures[16,17] or even by lumping protein demands by functionally related proteome groups. In contrast, the new version of the GECKO toolbox provides a systematic and robust parameterization procedure, leveraging the vastly accumulated knowledge of biochemistry research stored in public databases, ensuring the incorporation of biologically meaningful kinetic parameters even for poorly studied reactions and organisms.

The applicability of these other simple modeling formalisms to models for diverse species is limited as none of these methods has been provided as part of a generalized model-agnostic software implementation. Recently, a simplified variant of the MOMENT method (sMOMENT) was developed and embedded into an automated pipeline for generation and calibration of enzyme-constrained models of metabolism (AutoPACMEN)[70]. The pipeline was tested on the generation of an enzyme-constrained version of the *i*JO1366 metabolic reconstruction for *E. coli*, which also showed consistency with experimental data. This work represented a step forward in the field of constrain-based metabolic modeling, as it contributed to standardization of model generation and facilitating their utilization and applicability to other cases. However, due to the intrinsic trade-off between model simplicity and descriptive representation, a limitation of the sMOMENT method is its simplification of redundancies in metabolism, which just accounts for the

optimal way of catalyzing a given biochemical reaction, discarding the representation of alternative isoforms that might be relevant under certain conditions. In GECKO ecModels, all enzymes for which a gene-E.C. number relationship exists are included in the model structure. As traditional FBA simulations rely on optimality principles one could, in principle, expect the same predicted flux distributions by sMOMENT and GECKO ecModels. Nonetheless, the explicit incorporation of all enzymes in a metabolic network enables explanation of protein expression profiles that deviate from optimality in order to gain robustness to changes in the environment, as it has been recently shown by the integration of a regulatory nutrient-signaling Boolean network together with an ecModel for *S. cerevisiae*'s central carbon metabolism[71].

In conclusion, GECKO 2.0 together with the development of the automated pipeline *ecModels container* facilitates the generation, standardization, utilization, exchange and community development of ecModels through a transparent version-controlled environment. This tool provides a dynamic, and potentially increasing, catalog of updated ecModels trying to close the gap between model developers and final users and reduce the time-consuming tasks of model maintenance. We are confident that this will enable wide use of ecModels in basic science for obtaining novel insight into the function of metabolism, as well as in synthetic biology and metabolic engineering for design of strains with improved functionalities, e.g., for high-level production of valuable chemicals.

Methods

Automation pipeline and version-controlled hosting of the ecModels container. The ecModels repository is used to version-control the pipeline code and the resulting models. The pipeline is restricted to 2 short Python files, whose role is to decide when models need to be updated based on a configuration file config.ini, and to consequently invoke the use of GECKO for each model. Updates are deemed necessary when either the underlying dependencies (i.e., GECKO, RAVEN and COBRA toolboxes, the Gurobi solver, and libSMBL) or the source GEMs are independently updated to a new version (release) in their respective repositories.

The pipeline is designed be automatic and to not require supervision. It was developed to work with both version-controlled GEMs and GEMs downloadable from a URL, updating the version in the configuration after a new ecModel is obtained. For easy review, the pipeline log is publicly available under the *Actions* tab of the GitHub repository. The computation is performed through a self-hosted GitHub runner, further leveraging the transparent nature of the GitHub platform and the *git* version- control system. The resulting ecModel and updated configuration are committed to the repository, with the changes being made available for review through a pull request. Additionally, the GECKO output is also replicated in the pull request body. The *ecModels container* thus continues the transparency and reproducibility of the source models.

Quantification of absolute protein concentrations for *S. cerevisiae*, *Y. lipolytica* and *K. marxianus*. Total protein extraction for the strains *Saccharomyces cerevisiae* CEN.PK113-7D (standard, low pH, high temperature, osmotic stress), *Kluyveromyces marxianus* CBS6556 (standard, low pH, high temperature, osmotic stress) and *Yarrowia lipolytica* W29 (standard, low pH, high temperature) was conducted as described in the supplementary methods section in the Supplementary Information File 1. Three reference samples (hereafter, 'bulk' samples), one per strain, were constructed by pooling 5 µg of each experimental sample. Aliquots of 15 µg of total protein extract from each sample (3 strains x 4 conditions x 3 replicates) and the three bulks were separated on one- dimensional sodium dodecyl-sulfate–polyacrylamide gel electrophoresis short-migration gels (1 × 1 cm lanes, Invitrogen, NP321BOX). Yeast proteins digestion was performed on excised bands from gel gradient and digested peptides of UPS2 (Sigma) were used as external standards for absolute protein quantification (more details in the supplementary methods section in the Supplementary Information File 1). Four microliters of the different peptide mixtures (800 ng for yeast peptides and 949 ng for bulks) were analyzed using an Orbitrap Fusion™ Lumos™ Tribrid™ mass spectrometer (Thermo Fisher Scientific).

Protein identification was performed using the open-source search engine X! Tandem pipeline 3.4.4[72]. Data filtering was set to peptide *E*-value < 0.01 and protein log(*E*-value) < −3. Relative quantification of protein abundances was carried out using the Normalized Spectral Abundance Factor (NSAF)[73] and the NSAF values obtained from UPS2 proteins in bulk samples were used to determine the suitable regression curves that allowed the conversion from relative protein abundance into absolute terms. The regression curves parameters for protein abundance quantification are shown in the supplementary methods section in the Supplementary Information File 1.

Simulation of condition-dependent flux distributions. Simulation of cellular phenotypes for conditions of environmental stress at low dilution rates with GEMs were performed by first setting bounds on measured glucose uptake and byproduct secretion rates according to experimental data from previous studies on chemostats[64]. Then the biomass production rate was constrained (both upper and lower bounds) with the experimental dilution rate (0.1 h^{-1}). Maximization of the non-growth associated maintenance pseudo-reaction was set as an objective function for the parsimonious FBA problem as a representation of the additional energy demands for regulation of cellular growth at non-optimal conditions. The same procedure was followed for simulations with ecModels constrained by a total protein pool. For the case of ecModels with proteomics constraints, the same set of constraints was used but the objective function was set as minimization of the total usage of unmeasured proteins, assuming that the regulatory machinery for stress tolerance is represented by the condition-specific protein expression profile.

Prediction of microbial batch growth rates. Batch cellular growth was simulated by allowing unconstrained uptake of all nutrients present in minimal mineral media, enabling a specific carbon source uptake reaction for each case while blocking the rest of the uptake reactions and allowing unconstrained secretion rates for all exchangeable metabolites. Maximization of the biomass production rate was used as an objective function for the resulting FBA problem. For prediction of total protein demands on unlimited nutrient conditions, media constraints were set as expressed above and experimental batch growth rate values were fixed as both lower and upper bounds for the biomass production pseudo-reaction. The total protein pool exchange pseudo-reaction was then unconstrained and set as an objective function to minimize, assuming that when exposed to unlimited availability of nutrients the total mass of protein available for catalyzing metabolic reactions becomes the limiting resource for cells. The solveLP function, available in the RAVEN toolbox (v2.4.3), was used for solving all FBA problems in this study.

Reporting summary. Further information on research design is available in the Nature Research Reporting Summary linked to this article.

Data availability

Mass spectrometry raw data that support the findings of this study have been deposited in PRIDE database[74] with the dataset identifier PXD012836. The processed proteomics datasets are available in our GitHub repository at: https://github.com/SysBioChalmers/GECKO2_simulations/tree/v1.0.1/data/proteomics. All collected kinetic data for the study presented in Supplementary Information File are available at: https://github.com/SysBioChalmers/Enzyme-parameters-analysis/tree/master/data. The generated computational models used for this study are available at: https://github.com/SysBioChalmers/ecModels/tree/v1.0.0. Data for reproduction of all main and supplementary figures are provided in the Sournce Data: Data Source file 1, Data Source File 2, and Data Source File 3. Source data are provided with this paper.

Code availability

The source code of the updated GECKO toolbox is available at: https://github.com/SysBioChalmers/GECKO/releases/tag/v2.0.2[75]. The source code for ecModels container can be accessed at: https://github.com/SysBioChalmers/ecModels/tree/v1.0.0[76]. All custom scripts for simulations included in this study can be found at: https://github.com/SysBioChalmers/GECKO2_simulations/releases/tag/v1.0.1[77]. All the necessary scripts for reproducing the k_{cat} parameters analysis in the Supplementary Information File 1 are available at: https://github.com/SysBioChalmers/Enzyme-parameters-analysis/releases/tag/v1.0.0[78]. All of these repositories are public and open to collaborative continuous development.

Received: 21 March 2021; Accepted: 16 June 2022;
Published online: 30 June 2022

References

1. Lu, H. et al. A consensus *S. cerevisiae* metabolic model Yeast8 and its ecosystem for comprehensively probing cellular metabolism. *Nat. Commun.* **10**, 3586 (2019).
2. Robinson, J. L. et al. An atlas of human metabolism. *Sci. Signal.* **13**, eaaz1482 (2020).
3. Tiukova, I. A., Prigent, S., Nielsen, J., Sandgren, M. & Kerkhoven, E. J. Genome-scale model of Rhodotorula toruloides metabolism. *Biotechnol. Bioeng.* **116**, 3396–3408 (2019).
4. Monk, J. M. et al. iML1515, a knowledgebase that computes *Escherichia coli* traits. *Nat. Biotechnol.* **35**, 904–908 (2017).

5. Kerkhoven, E. J., Pomraning, K. R., Baker, S. E. & Nielsen, J. Regulation of amino-acid metabolism controls flux to lipid accumulation in Yarrowia lipolytica. *NPJ Syst. Biol. Appl.* **2**, 16005 (2016).
6. Marcišauskas, S., Ji, B. & Nielsen, J. Reconstruction and analysis of a Kluyveromyces marxianus genome-scale metabolic model. *BMC Bioinforma.* **20**, 551 (2019).
7. Meadows, A. L. et al. Rewriting yeast central carbon metabolism for industrial isoprenoid production. *Nature* **537**, 694–697 (2016).
8. Chen, X. et al. Fumaric acid production by Torulopsis glabrata: Engineering the urea cycle and the purine nucleotide cycle. *Biotechnol. Bioeng.* **112**, 156–167 (2015).
9. Mishra, P. et al. Genome-scale model-driven strain design for dicarboxylic acid production in Yarrowia lipolytica. *BMC Syst. Biol.* **12**, 12 (2018).
10. Agren, R. et al. Identification of anticancer drugs for hepatocellular carcinoma through personalized genome-scale metabolic modeling. *Mol. Syst. Biol.* **10**, 721 (2014).
11. Mardinoglu, A. et al. Genome-scale metabolic modelling of hepatocytes reveals serine deficiency in patients with non-alcoholic fatty liver disease. *Nat. Commun.* **5**, 3083 (2014).
12. Gatto, F., Miess, H., Schulze, A. & Nielsen, J. Flux balance analysis predicts essential genes in clear cell renal cell carcinoma metabolism. *Sci. Rep.* **5**, 10738 (2015).
13. Orth, J. D., Thiele, I. & Palsson, B. Ø. What is flux balance analysis? *Nat. Biotechnol.* **28**, 245–248 (2010).
14. Ibarra, R. U., Edwards, J. S. & Palsson, B. O. *Escherichia coli* K-12 undergoes adaptive evolution to achieve in silico predicted optimal growth. *Nature* **420**, 186–189 (2002).
15. Mahadevan, R. & Schilling, C. H. The effects of alternate optimal solutions in constraint-based genome-scale metabolic models. *Metab. Eng.* **5**, 264–276 (2003).
16. Beg, Q. K. et al. Intracellular crowding defines the mode and sequence of substrate uptake by *Escherichia coli* and constrains its metabolic activity. *Proc. Natl Acad. Sci. USA* **104**, 12663–12668 (2007).
17. Vazquez, A. et al. Impact of the solvent capacity constraint on E. coli metabolism. *BMC Syst. Biol.* **2**, 7 (2008).
18. Molenaar, D., van Berlo, R., de Ridder, D. & Teusink, B. Shifts in growth strategies reflect tradeoffs in cellular economics. *Mol. Syst. Biol.* **5**, 323 (2009).
19. Zhuang, K., Vemuri, G. N. & Mahadevan, R. Economics of membrane occupancy and respiro-fermentation. *Mol. Syst. Biol.* **7**, 500 (2011).
20. Shlomi, T., Benyamini, T., Gottlieb, E., Sharan, R. & Ruppin, E. Genome-scale metabolic modeling elucidates the role of proliferative adaptation in causing the Warburg effect. *PLoS Comput. Biol.* **7**, e1002018 (2011).
21. Mori, M., Hwa, T., Martin, O. C., De Martino, A. & Marinari, E. Constrained allocation flux balance analysis. *PLOS Comput. Biol.* **12**, e1004913 (2016).
22. Nilsson, A. & Nielsen, J. Metabolic trade-offs in yeast are caused by F1F0-ATP synthase. *Sci. Rep.* **6**, 22264 (2016).
23. Adadi, R., Volkmer, B., Milo, R., Heinemann, M. & Shlomi, T. Prediction of microbial growth rate versus biomass yield by a metabolic network with kinetic parameters. *PLoS Comput. Biol.* **8**, e1002575 (2012).
24. Nilsson, A., Björnson, E., Flockhart, M., Larsen, F. J. & Nielsen, J. Complex I is bypassed during high intensity exercise. *Nat. Commun.* **10**, 5072 (2019).
25. Chen, Y. & Nielsen, J. Energy metabolism controls phenotypes by protein efficiency and allocation. *Proc. Natl Acad. Sci. USA* **116**, 17592–17597 (2019).
26. van Hoek, M. J. A. & Merks, R. M. H. Redox balance is key to explaining full vs. partial switching to low-yield metabolism. *BMC Syst. Biol.* **6**, 22 (2012).
27. van Hoek, M. J. & Merks, R. M. Redox balance is key to explaining full vs. partial switching to low-yield metabolism. *BMC Syst. Biol.* **6**, 22 (2012).
28. Jeske, L., Placzek, S., Schomburg, I., Chang, A. & Schomburg, D. BRENDA in 2019: a European ELIXIR core data resource. *Nucleic Acids Res.* **47**, D542–D549 (2019).
29. Wittig, U., Rey, M., Weidemann, A., Kania, R. & Müller, W. SABIO-RK: an updated resource for manually curated biochemical reaction kinetics. *Nucleic Acids Res.* **46**, D656–D660 (2018).
30. O'Brien, E. J., Lerman, J. A., Chang, R. L., Hyduke, D. R. & Palsson, B. Ø. Genome-scale models of metabolism and gene expression extend and refine growth phenotype prediction. *Mol. Syst. Biol.* **9**, 693 (2013).
31. O'Brien, E. J. & Palsson, B. O. Computing the functional proteome: recent progress and future prospects for genome-scale models. *Curr. Opin. Biotechnol.* **34**, 125–134 (2015).
32. Yang, L. et al. Principles of proteome allocation are revealed using proteomic data and genome-scale models. *Sci. Rep.* **6**, 36734 (2016).
33. King, Z. A., O'Brien, E. J., Feist, A. M. & Palsson, B. O. Literature mining supports a next-generation modeling approach to predict cellular byproduct secretion. *Metab. Eng.* **39**, 220–227 (2017).
34. Lerman, J. A. et al. In silico method for modelling metabolism and gene product expression at genome scale. *Nat. Commun.* **3**, 929 (2012).
35. Chen, Y. et al. Proteome constraints reveal targets for improving microbial fitness in nutrient-rich environments. *Mol. Syst. Biol.* **17**, (2021).
36. Bulović, A. et al. Automated generation of bacterial resource allocation models. *Metab. Eng.* https://doi.org/10.1016/j.ymben.2019.06.001 e10093 (2019).
37. Goelzer, A. et al. Quantitative prediction of genome-wide resource allocation in bacteria. *Metab. Eng.* **32**, 232–243 (2015).
38. Sánchez, B. J. et al. Improving the phenotype predictions of a yeast genome-scale metabolic model by incorporating enzymatic constraints. *Mol. Syst. Biol.* **13**, 935 (2017).
39. Li, G. et al. Bayesian genome scale modelling identifies thermal determinants of yeast metabolism. *Nat. Commun.* **12**, 190 (2021).
40. Massaiu, I. et al. Integration of enzymatic data in Bacillus subtilis genome-scale metabolic model improves phenotype predictions and enables in silico design of poly-γ-glutamic acid production strains. *Microb. Cell Fact.* **18**, 3 (2019).
41. Ye, C. et al. Improving lysine production through construction of an Escherichia coli enzyme-constrained model. *Biotechnol. Bioeng.* **117**, 3533–3544 (2020).
42. Chen, Y. et al. Genome-scale modeling for Bacillus coagulans to understand the metabolic characteristics. *Biotechnol. Bioeng.* **117**, 3545–3558 (2020).
43. Sulheim, S. et al. Enzyme-constrained models and omics analysis of *Streptomyces coelicolor* reveal metabolic changes that enhance heterologous production. *iScience* **23**, 101525 (2020).
44. Heirendt, L. et al. Creation and analysis of biochemical constraint-based models using the COBRA Toolbox v.3.0. *Nat. Protoc.* **14**, 639–702 (2019).
45. Ebrahim, A., Lerman, J. A., Palsson, B. O. & Hyduke, D. R. COBRApy: constraints-based reconstruction and analysis for python. *BMC Syst. Biol.* **7**, 74 (2013).
46. Olivier, B. G. & Bergmann, F. T. The systems biology markup language (SBML) level 3 package: flux balance constraints. *J. Integr. Bioinform.* **12**, 269 (2015).
47. Placzek, S. et al. BRENDA in 2017: new perspectives and new tools in BRENDA. *Nucleic Acids Res.* **45**, D380–D388 (2017).
48. Kanehisa, M., Furumichi, M., Sato, Y., Ishiguro-Watanabe, M. & Tanabe, M. KEGG: integrating viruses and cellular organisms. *Nucleic Acids Res.* **49**, D545–D551 (2021).
49. Bar-Even, A. et al. The moderately efficient enzyme: evolutionary and physicochemical trends shaping enzyme parameters. *Biochemistry* **50**, 4402–4410 (2011).
50. Wang, H. et al. RAVEN 2.0: a versatile toolbox for metabolic network reconstruction and a case study on Streptomyces coelicolor. *PLOS Comput. Biol.* **14**, e1006541 (2018).
51. Ochoa-Estopier, A. & Guillouet, S. E. D-stat culture for studying the metabolic shifts from oxidative metabolism to lipid accumulation and citric acid production in Yarrowia lipolytica. *J. Biotechnol.* **170**, 35–41 (2014).
52. Aggelis, G. & Komaitis, M. Enhancement of single cell oil production by Yarrowia lipolytica growing in the presence of Teucrium polium L. aqueous extract. *Biotechnol. Lett.* https://doi.org/10.1023/A:1005591127592 (1999).
53. Overkamp, K. M. et al. In vivo analysis of the mechanisms for oxidation of cytosolic NADH by *Saccharomyces cerevisiae* mitochondria. *J. Bacteriol.* **182**, 2823–2830 (2000).
54. Li, M. & Borodina, I. Application of synthetic biology for production of chemicals in yeast Saccharomyces cerevisiae. *FEMS Yeast Res.* **15**, n/a–n/a (2014).
55. Rocha, I. et al. OptFlux: an open-source software platform for in silico metabolic engineering. *BMC Syst. Biol.* **4**, 45 (2010).
56. Choi, H. S., Lee, S. Y., Kim, T. Y. & Woo, H. M. In silico identification of gene amplification targets for improvement of lycopene production. *Appl. Environ. Microbiol.* **76**, 3097–3105 (2010).
57. Norsigian, C. J. et al. BiGG Models 2020: multi-strain genome-scale models and expansion across the phylogenetic tree. *Nucleic Acids Res.* **48**, D402–D406 (2019).
58. Björkeroth, J. et al. Proteome reallocation from amino acid biosynthesis to ribosomes enables yeast to grow faster in rich media. *Proc. Natl Acad. Sci. USA* **117**, 21804–21812 (2020).
59. Yu, R. et al. Nitrogen limitation reveals large reserves in metabolic and translational capacities of yeast. *Nat. Commun.* **11**, 1881 (2020).
60. Campbell, K. et al. Building blocks are synthesized on demand during the yeast cell cycle. *Proc. Natl Acad. Sci. USA* **117**, 7575–7583 (2020).
61. Lahtvee, P.-J. et al. Absolute quantification of protein and mRNA abundances demonstrate variability in gene-specific translation efficiency in Yeast. *Cell Syst.* **4**, 495–504.e5 (2017).
62. Di Bartolomeo, F. et al. Absolute yeast mitochondrial proteome quantification reveals trade-off between biosynthesis and energy generation during diauxic shift. *Proc. Natl Acad. Sci. USA* **117**, 7524–7535 (2020).

63. Sánchez, B. J. et al. Benchmarking accuracy and precision of intensity-based absolute quantification of protein abundances in Saccharomyces cerevisiae. *Proteomics* **21**, 2000093 (2021).
64. Doughty, T. W. et al. Stress-induced expression is enriched for evolutionarily young genes in diverse budding yeasts. *Nat. Commun.* **11**, 2144 (2020).
65. Kauffman, K. J., Prakash, P. & Edwards, J. S. Advances in flux balance analysis. *Curr. Opin. Biotechnol.* **14**, 491–496 (2003).
66. Teusink, B. et al. Analysis of growth of Lactobacillus plantarum WCFS1 on a complex medium using a genome-scale metabolic model. *J. Biol. Chem.* **281**, 40041–40048 (2006).
67. Beard, D. A., Liang, S. & Qian, H. Energy BAlance for Analysis of Complex Metabolic Networks. *Biophys. J.* **83**, 79–86 (2002).
68. Maurice Cheung, C. Y., George Ratcliffe, R. & Sweetlove, L. J. A method of accounting for enzyme costs in flux balance analysis reveals alternative pathways and metabolite stores in an illuminated arabidopsis leaf. *Plant Physiol.* https://doi.org/10.1104/pp.15.00880 (2015).
69. Patil, K. R. & Nielsen, J. Uncovering transcriptional regulation of metabolism by using metabolic network topology. *Proc. Natl Acad. Sci. USA* **102**, 2685–2689 (2005).
70. Bekiaris, P. S. & Klamt, S. Automatic construction of metabolic models with enzyme constraints. *BMC Bioinforma.* **21**, 19 (2020).
71. Österberg, L. et al. A novel yeast hybrid modeling framework integrating Boolean and enzyme-constrained networks enables exploration of the interplay between signaling and metabolism. *PLOS Comput. Biol.* **17**, e1008891 (2021).
72. Langella, O. et al. X!TandemPipeline: a tool to manage sequence redundancy for protein inference and phosphosite identification. *J. Proteome Res.* **16**, 494–503 (2017).
73. Zybailov, B. et al. Statistical analysis of membrane proteome expression changes in *Saccharomyces cerevisiae*. *J. Proteome Res.* **5**, 2339–2347 (2006).
74. Vizcaíno, J. A. et al. 2016 update of the PRIDE database and its related tools. *Nucleic Acids Res.* **44**, D447–D456 (2016).
75. Domenzain, I., Sánchez, B. J., Kerkhoven, E. J. Reconstruction of a catalogue of genome-scale metabolic models with enzymatic constraints using GECKO 2.0. Repository name: GECKO. https://doi.org/10.5281/zenodo.6631788 (2022).
76. Domenzain, I., Sánchez, B. J., Anton, M. Reconstruction of a catalogue of genome-scale metabolic models with enzymatic constraints using GECKO 2.0. Repository name: ecModels. https://doi.org/10.5281/zenodo.6631421 (2022).
77. Domenzain, I. Reconstruction of a catalogue of genome-scale metabolic models with enzymatic constraints using GECKO 2.0. Repository name: GECKO2 simulations. https://doi.org/10.5281/zenodo.6628822 (2022).
78. Domenzain, I. Reconstruction of a catalogue of genome-scale metabolic models with enzymatic constraints using GECKO 2.0. Repository name: enzyme parameters analysis. https://doi.org/10.5281/zenodo.6624399 (2022).

Acknowledgements

We are grateful to Feiran Li, Raphaël Ferreira, Jonathan Robinson, and all the GECKO users that have provided feedback for improving our toolbox and extending its range of applications and to the CHASSY project consortium for having motivated and supported this work. This project has received funding from the European Union's Horizon 2020 Framework Programme for Research and Innovation—Grant Agreements No. 720824 to I.D., A.M.O., C.H., V.S., and J.P.M. 686070 to B.S. and 760798 to M.A. This work was also supported by the Knut and Alice Wallenberg Foundation and The Novo Nordisk Foundation—Grant no. NNF10CC1016517 to J.N.

Author contributions

Conceptualization: I.D., B.S., M.A., E.J.K., and J.N.; data curation: A.M.O., and C.H.; formal analysis: I.D.; funding acquisition: J.N.; methodology: I.D.; project administration: J.N.; software: I.D., B.S., and M.A.; supervision: V.S., J.P.M., N.S., and J.N.; validation: I.D.; visualization: I.D.; writing—original draft: I.D., B.S., M.A, E.J.K.; writing—review and editing: I.D., M.A., V.S., J.P.M., N.S., and J.N.

Funding

Open access funding provided by Chalmers University of Technology.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41467-022-31421-1.

Correspondence and requests for materials should be addressed to Jens Nielsen.

Peer review information *Nature Communications* thanks Priyanka Baloni and the other, anonymous, reviewer(s) for their contribution to the peer review of this work. Peer reviewer reports are available.

Reprints and permission information is available at http://www.nature.com/reprints

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Paper IV:

Genome-scale modeling drives 70-fold improvement of intracellular heme production in Saccharomyces cerevisiae

Ishchuk, O. P., Domenzain, I., Sánchez, B. J., Muñiz-Paredes, F., Martínez, J. L., Nielsen, J., & Petranovic, D.

Proceedings of the National Academy of Sciences, 2022

PNAS RESEARCH ARTICLE | SYSTEMS BIOLOGY OPEN ACCESS

Genome-scale modeling drives 70-fold improvement of intracellular heme production in *Saccharomyces cerevisiae*

Olena P. Ishchuk[a,1], Iván Domenzain[a,b], Benjamín J. Sánchez[a,b,c,d], Facundo Muñiz-Paredes[a], José L. Martínez[a,c], Jens Nielsen[a,b,e], and Dina Petranovic[a,b,1]

Edited by Costas Maranas, The Pennsylvania State University, University Park, PA; received May 1, 2021; accepted June 7, 2022 by Editorial Board Member Stephen J. Benkovic

Heme is an oxygen carrier and a cofactor of both industrial enzymes and food additives. The intracellular level of free heme is low, which limits the synthesis of heme proteins. Therefore, increasing heme synthesis allows an increased production of heme proteins. Using the genome-scale metabolic model (GEM) Yeast8 for the yeast *Saccharomyces cerevisiae*, we identified fluxes potentially important to heme synthesis. With this model, *in silico* simulations highlighted 84 gene targets for balancing biomass and increasing heme production. Of those identified, 76 genes were individually deleted or overexpressed in experiments. Empirically, 40 genes individually increased heme production (up to threefold). Heme was increased by modifying target genes, which not only included the genes involved in heme biosynthesis, but also those involved in glycolysis, pyruvate, Fe-S clusters, glycine, and succinyl-coenzyme A (CoA) metabolism. Next, we developed an algorithmic method for predicting an optimal combination of these genes by using the enzyme-constrained extension of the Yeast8 model, ecYeast8. The computationally identified combination for enhanced heme production was evaluated using the heme ligand-binding biosensor (Heme-LBB). The positive targets were combined using CRISPR-Cas9 in the yeast strain (IMX581-*HEM15-HEM14-HEM3-*Δ*shm1-HEM2-*Δ*hmx1-FET4-*Δ*gcv2-HEM1-*Δ*gcv1-HEM13*), which produces 70-fold-higher levels of intracellular heme.

genome-scale modeling | heme | *Saccharomyces cerevisiae* | metabolic engineering | heme ligand-binding biosensor

Significance

Heme availability in the cell enables the proper folding and function of enzymes, which carry heme as a cofactor. Using genome-scale modeling, we identified metabolic fluxes and genes that limit heme production. Our study experimentally validates ecYeast8 model predictions. Moreover, we developed an approach to predict gene combinations, which provides an *in silico* design of a viable strain able to overproduce the metabolite of interest. Using our approach, we constructed a yeast strain that produces 70-fold-higher levels of intracellular heme. With its high-capacity metabolic subnetwork, our engineered strain is a suitable platform for the production of additional heme enzymes. The heme ligand-binding biosensor (Heme-LBB) detects the cotranslational incorporation of heme into the heme-protein hemoglobin.

Heme is a cofactor of essential enzymes for aerobic life within the three domains of life (archaea, bacteria, and eukarya). The heme molecule consists of a porphyrin ring that surrounds an iron atom, which alternates between its ferric and ferrous states in the oxidation and reduction reactions. Heme-containing proteins (HCPs) have several functions. For example, HCPs transport electrons in the respiratory chain in mitochondria and are crucial for energy production, transport molecular oxygen in globin proteins (e.g., hemoglobin in humans), and protect cells from oxidative damage (1–4). The heme biosynthetic pathway is conserved and tightly regulated to supply heme at levels to meet cellular demands. The cotranslational incorporation of heme into heme proteins governs their folding process (5, 6). The intracellular availability of heme is crucial for the production of heme proteins, which denature and lose their function without heme.

Because of their biological importance, heme and HCPs are a central topic in molecular cell biology, with basic research occurring together with applications in medicine and technology. The production of heme and heme proteins has been a focus of research in microbial metabolic engineering. For example, research on blood substitutes focuses on human hemoglobin (7, 8), and plant-derived hemoglobin provides vegetarian protein (artificial meat with a lower carbon footprint) (9). Heme was used to improve charging of lithium batteries (10) and in the bioremediation of sulfite waste (11). Cytochromes and their new mutant forms catalyze novel chemical reactions with silicon (12) and were evolved to perform novel chemical reactions (13). The heterologous production of heme proteins is, however, challenging due to the limited amount of free heme and the complexity of the metabolic network in the cell.

While a heme-biosynthesis pathway is conserved in nature, the precursor 5-aminolevulinic acid (5-ALA) is synthesized distinctly in different organisms. In the C4 pathway, the precursor 5-ALA is produced from glycine and from succinyl-coenzyme A (CoA) (the C4 intermediate of the tricarboxylic acid [TCA] cycle) in yeast, birds, mammals, and purple nonsulfur photosynthetic bacteria. In contrast, in the C5 pathway, the precursor 5-ALA is produced from alpha-ketoglutarate (the C5 intermediate of the TCA cycle) in algae, plants, and bacteria such as *Escherichia coli* (14).

Author contributions: O.P.I. and D.P. designed research; O.P.I., I.D., B.J.S., and F.M.-P. performed research; I.D., O.P.I., and B.J.S. contributed new reagents/analytic tools; O.P.I., I.D., B.J.S., F.M.-P., J.L.M., J.N., and D.P. analyzed data; and O.P.I., I.D., and B.J.S. wrote the paper.

The authors declare no competing interest.

This article is a PNAS Direct Submission. C.M. is a guest editor invited by the Editorial Board.

[1]To whom correspondence may be addressed. Email: ishchuk@chalmers.se or dina.petranovic@chalmers.se

This article contains supporting information online at http://www.pnas.org/lookup/suppl/doi:10.1073/pnas.2108245119/-/DCSupplemental.

Published July 18, 2022.

In *E. coli*, heme production has been increased by metabolic engineering of the pathways for 5-ALA synthesis, both native (C5) and heterologous (C4). The metabolic engineering studies using the C4 pathway increased heme production by overexpression of the *Rhodobacter sphaeroides hemA* gene (encoding ALA synthase), which produces the 5-ALA precursor; by overexpressing the native *coaA* gene (encoding pantothenate kinase), which produces CoA; and by overexpression of genes for heme biosynthesis. This engineering strategy yielded 3.3 μmol/L (15) and 9.1 $\mu mol/g_{cell}$ (16) of heme. By overexpressing genes for heme production via the C5 pathway and by deleting genes of competing pathways, 51.5 mg/L total heme was produced (17). In the same strain, metabolic engineering of a heme-secretory pathway and feed-control optimization of substrates in fed-batch cultivation increased the production of total heme to 239 mg/L (17).

In the unicellular eukaryote and established yeast cell-factory *Saccharomyces cerevisiae*, heme is synthesized through the C4 pathway (14). To improve the production of heme and heme proteins in *S. cerevisiae*, metabolic engineering studies have overexpressed genes encoding the known rate-limiting enzymes for heme biosynthesis (18–21) and have engineered oxygen sensing involved in heme biosynthesis regulation (22). To increase the production of the first intermediate of the heme pathway, 5-ALA, the *HEM1* and *ACO2* genes were overexpressed (23). However, the contribution of overall metabolism to heme production has not been analyzed.

The impressive development of heme production in *E. coli*, however, has had some limitations, such as weak tolerance to acidic pH and phage sensitivity. As *E. coli* produces endotoxins, it is difficult to use *E. coli* directly in food production. In contrast, the *S. cerevisiae* yeast has greater tolerance for acidic pH and has been used for food production for millennia.

The *S. cerevisiae* has been analyzed with genome-scale metabolic models (GEMs) (24). For *S. cerevisiae,* GEM analysis has guided the construction of strains with optimized yields of industrial molecules (e.g., bioethanol, sesquiterpenes, vanillin, 2,3-butanediol, fumaric acid, succinate, amorphadiene, 3-hydroxypropionate, β-farnesene, and dihydroxyacetone phosphate [DHAP]) (24, 25). The measurement of metabolic compounds in screening has facilitated the development of new biosensors that can be used for novel applications in other organisms (26).

For *S. cerevisiae*, the consensus GEM (version 7.6) informed the engineering of strains with increased production of acetyl-CoA and malonyl-CoA in 2019 (27). The updated consensus Yeast8 model was followed by ecYeast8, which has additional constraints on the metabolic fluxes, representing enzymatic abundances. Enzyme-constrained GEMs improved the prediction of specific phenotypes (28, 29).

Studies of heme production have explored the modification of genes and their expression, improving our knowledge of particular pathways. Using metabolic GEMs to maximize the production of heme is the focus of this study. We used the 2019 enzyme-constrained ecYeast8 (29) to identify metabolic fluxes that are important for heme biosynthesis. Our systems-biology analysis and modification of the gene expression guided the optimization of a heme strain with 58 genes *in silico*. The sequential strain engineering increased intracellular heme production 70-fold. In optimization of sequentially accumulated gene modifications, we developed a heme biosensor, which detects heme availability and the incorporation of heme into hemoglobin protein. This heme ligand-binding biosensor (Heme-LBB), like earlier genetically encoded ratiometric heme sensors (30), is likely useful for heme detection in other organisms.

Our results are striking in terms of the dramatic increase in heme production and as a showcase of model-assisted synthetic biology. More importantly, our case study is one of the most rigorous in terms of evaluation of model-predicted targets for the widely used cell factory *S. cerevisiae*. As several of the model-predicted targets resulted in improved production, our paper represents a significant milestone in terms of a wider use of model-based design of yeast cell factories.

Results

Yeast8 Simulations of Metabolic Fluxes Impacting Heme Production. As an initial screening, we quantified the fluxes impacting heme production using flux balance analysis (FBA) tools available for *S. cerevisiae* at the start of our study. Using Yeast8 (29), we computed the theoretical biomass yield on glucose to be 0.1168 gDW/g for batch cultures, which is very close to the experimentally validated value of our strain (0.122 gDW/g). We followed a published approach (27, 31), which is the adaptation of the flux scanning based on enforced objective flux (FSEOF) method (32). To simulate physiologically relevant conditions and analyze heme production at suboptimal growth yields, we ran several simulations on glucose as the single carbon source, varying the biomass yield from half of the experimental yield to twice the value (Fig. 1). In each simulation, the objective function was to maximize heme production, computing for each biomass-yield condition an optimum solution. In these simulation-generated optimal solutions, the number of active fluxes was reduced by

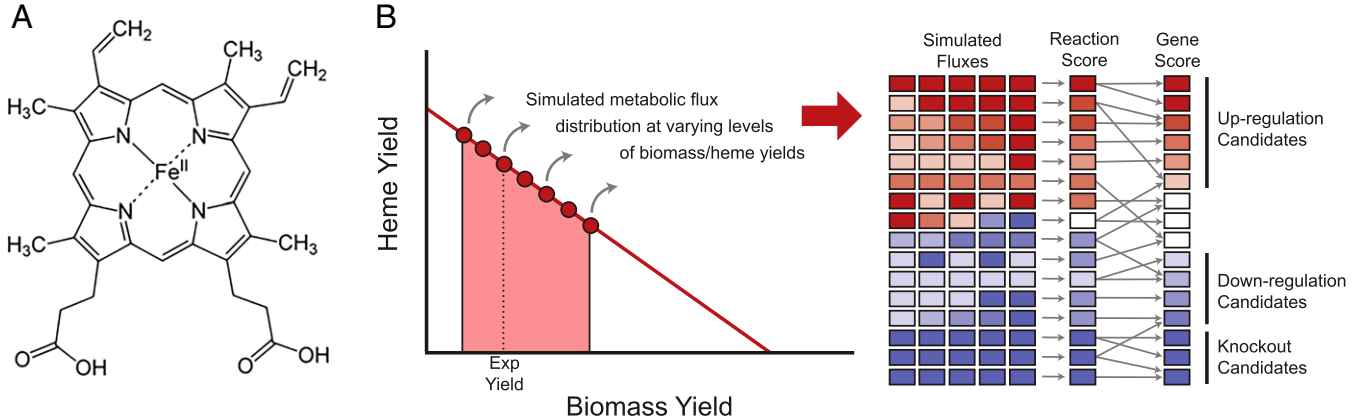


Fig. 1. The Yeast8 genome-scale model was used to find fluxes important for heme production to enable the construction of a heme yeast cell factory. (*A*) The structure of heme *b*, which is protoporphyrin IX with ferrous iron. (*B*) Simulations of heme production using *S. cerevisiae* Yeast8.0.1 model.

parsimonious FBA (33). From these simulations, scores were computed for each metabolic reaction in the network to detect which fluxes were consistently either increasing or decreasing as the biomass requirements decreased, an established strategy (27, 31, 32). Finally, using known reaction-gene associations, we converted those flux scores to gene scores, which indicate whether a gene has a monotonic behavior—that is, the flux scores selected genes that are consistently upregulated (score > 1), downregulated (0 < score < 1), or completely silenced (score = 0) (Fig. 1). This scoring predicted that 84 genes had a monotonic effect, including 62 genes being overexpressed and 8 genes deleted. Additionally, 14 genes were downregulated (among them, 6 were essential or required additional growth supplements when deleted: *OLE1*, *FAS1*, *FAS2*, *RNR2*, *CDS1*, and *CHO1*) (Dataset S1).

Validation of Individual Gene Targets Predicted by Yeast8. The gene targets predicted by Yeast8 were experimentally tested for their impact on heme production by modifying genes one at a time and by measuring intracellular heme concentration. As gene downregulation requires more fine tuning (e.g., by promoter modifications or gene silencing approaches), we tested the effect of gene deletions first by using the deletion strains from the yeast knockout (YKO) collection (34). We analyzed 16 *S. cerevisiae* BY4741 strains carrying single gene deletions: 8 strains from the downregulation group (*Δrnr1*, *Δrnr3*, *Δrnr4*, *Δcho2*, *Δopi3*, *Δpsd1*, *Δgpt2*, *Δale1*) and 8 deletion strains from the deletion group (*Δshm1*, *Δslc1*, *Δpro1*, *Δpro2*, *Δsfc1*, *Δyhm2*, *Δidh1*, *Δidh2*) (Fig. 2*A* and Dataset S1). The *Δrnr3* strain grew poorly and was excluded from further experimental analysis. The heme production of 15 strains was measured in two biological replicates after 24 h of cultivation in yeast extract-peptone-dextrose (YPD) medium (Fig. 2*A*).

Deletion of five out of the seven tested genes in the downregulation group (*OPI3* [encoding methylene-fatty-acyl-phospholipid synthase], *CHO2* [encoding phosphatidylethanolamine methyltransferase], *RNR1* [encoding major isoform of large subunit of ribonucleotide-diphosphate reductase], *RNR4* [encoding ribonucleotide-diphosphate reductase small subunit], and *ALE1*) validated the model predictions and increased heme production up to 70% compared to the BY4741 control strain (Fig. 2*A*). The deletion of two genes, *GPT2* (encoding glycerol-3-phosphate/DHAP sn-1 acyltransferase) and *PSD1* (encoding phosphatidylserine decarboxylase of the mitochondrial inner membrane), decreased heme production ∼10 to 50% compared to the BY4741 control strain (Fig. 2*A*).

Deletion of three out of the eight genes (identified to be deleted by Yeast8) increased heme production (Fig. 2*A*). The deletion of *SHM1* (encoding mitochondrial serine hydroxylmethyltransferase) resulted in a ∼11.5% increase in heme production, the deletion of the *ALE1* gene (encoding broad-specificity lysophospholipid acyltransferase) resulted in a ∼13% increase, and the deletion of *SFC1* (encoding mitochondrial succinate-fumarate transporter) resulted only in a ∼4% increase (Fig. 2*A*). On the other hand, the deletion of *SLC1* (encoding 1-acyl-sn-glycerol-3-phosphate acyltransferase) and *YHM2* (encoding citrate and oxoglutarate carrier protein), did not result in a significant increase in heme production compared with BY4741 (Fig. 2*A* and Dataset S1). Deletions of *PRO1* (encoding gamma-glutamyl kinase), *PRO2* (gamma-glutamyl phosphate reductase), and *IDH1* and *IDH2* (encoding subunits of mitochondrial NAD [+]-dependent isocitrate dehydrogenase) genes decreased heme production, contrary to the model predictions (Fig. 2*A* and Dataset S1). Both *PRO1* and *PRO2* gene deletions resulted in proline auxotrophy, and the resulting strains grew poorly in YPD. In summary, among the 15 tested gene candidates identified to be deleted or downregulated, 8 genes increased heme production.

We evaluated the overexpression of 61 of the 62 model gene targets (we could not amplify the *HMG2* gene) in the *S. cerevisiae* CEN.PK.113–11c strain background (Fig. 2*B*). For this purpose, we cloned the open reading frames (ORFs) of the 61 genes into the centromeric expression plasmid pRS316+prTEF1-terADH1 under control of strong constitutive promoter *TEF1*. Two transformants with expression cassettes for each of the 61 model target genes (predicted to be overexpressed) were used to evaluate heme production (Fig. 2*B* and Dataset S1). The highest heme production (∼300% average increase) was observed upon the overexpression of the *HEM13* (encoding coproporphyrinogen III oxidase) heme biosynthetic gene (Fig. 2*B*). Under normal conditions, the *HEM13* is transcriptionally repressed by Rox1 (22, 35), and expressing it under the promoter *TEF1* will increase the protein abundance independent of the oxygen and heme levels. Overexpressing other heme biosynthetic genes—such as *HEM14* (encoding protoporphyrinogen oxidase), *HEM2* (encoding aminolevulinate dehydratase), *HEM15* (encoding ferrochelatase), *HEM3* (encoding porphobilinogen deaminase), and *HEM12* (encoding uroporphyrinogen decarboxylase)—also increased heme production from ∼20 to 70%, respectively (Fig. 2 *B* and *C*). The *HEM2*, *HEM3*, and *HEM12* genes have been reported to be rate-limiting steps in heme biosynthesis (18–20). Overexpression of *HEM1* (encoding 5-aminolevulinate synthase) did not improve heme production at 48 h of fermentation (Fig. 2*B*), and the overexpression of *HEM4* (encoding uroporphyrinogen III synthase) resulted in substantially reduced yeast growth. We speculate this was caused by the accumulation of uroporphyrinogen III, which is toxic when oxidized (36). In addition to heme-biosynthetic genes, the overexpression of single genes involved in iron homeostasis and Fe-S cluster proteins (*YAH1* and *ARH1*), glutamate biosynthesis (*GLT1*), pyruvate metabolism and its transport (*PYC1*, *PYC2*, *MPC1*, *MPC2*, *MPC3*), fumarate reductase (*FRD1*), malate dehydrogenase (*MDH2*), glycolysis (*PFK1*, *PFK2*, *TDH1*, *TDH2, TDH3*), amino acids, iron, protons, and water transport (*AGC1*, *FET4*, *FET3*, *PMA1*, *PMA2*, *AQY1*, and *AQY2*) increased heme production up to 40% compared to the control strain carrying the empty vector pRS316 (Fig. 2 *B* and *C*). In summary, among the 61 overexpression targets tested, 32 increased heme production (Fig. 2*C*), which is a 52% success rate of model predictions.

Refining the Simulations of Heme Production Fluxes Using ecYeast8. We used the enzyme-constrained version of the Yeast8 model (ecYeast8) to refine model simulations and to evaluate the combinatorial effects of the gene targets (Fig. 3*A* and *SI Appendix*). The ecYeast8 model accounts for the activity of metabolic enzymes as constraints on the reactions in the network. These constraints are limited by the total amount of available protein mass, yielding a drastic reduction of the variability of the metabolic fluxes and notable improvements on phenotype predictions for *S. cerevisiae*'s metabolism (28, 29). Simulations for optimization of heme production using ecYeast8 were performed following the same procedure as with the Yeast8 model; in this case, candidate gene targets for downregulation (0 < gene score ≤ 1) were discarded. Additionally, as enzyme-constrained models enable a direct assessment of the effects of enzyme activity perturbations, the enzyme usage variability analysis and mechanistic simulations for the individual gene modifications were implemented for heme production (Fig. 3*B* and *SI Appendix*). This allowed the prediction of 80 gene targets (Dataset S2) by ecYeast8. Comparing the target

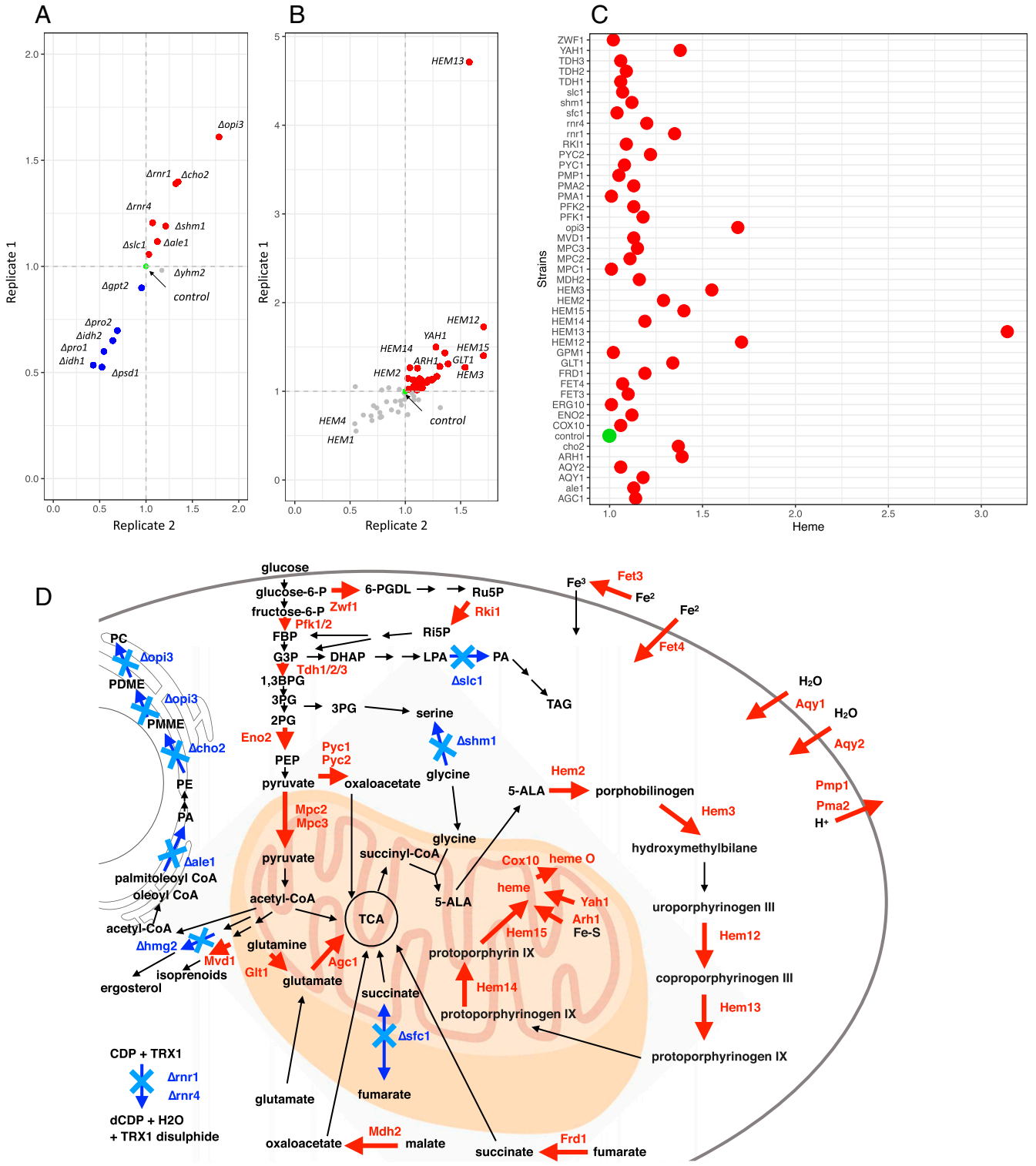


Fig. 2. Experimental validation of Yeast8 gene targets. (*A*) Heme production (fold-change) of 15 gene KO strains from the YKO collection (BY4741 strain background). BY4741 strain served as a control to normalize data (shown in green). Two replicates were used in the analysis. Heme was extracted from eight OD_{600} of cells. The gene targets where heme production was higher than the control are highlighted in red. The gene targets where heme production was lower than the control are highlighted in blue. (*B*) Heme production (fold-change) of strains carrying 61 model genes overexpressed under control of the *TEF1* promoter using a centromeric plasmid in CEN.PK113-11C strain background. Heme was extracted from eight OD_{600} of cells. CEN.PK113-11C carrying empty vector served as a control to normalize data (shown in green). Two replicates were used in the analysis. (*C*) Heme production of strains with gene modifications that improved heme production the most. Average value of two replicates was used. Heme was extracted from eight OD_{600} of cells. (*D*) Schematic overview of metabolism with Yeast8 targets, which experimentally improved heme production.

lists of both models (Datasets S1 and S2), 40 genes were found to overlap between Yeast8 and ecYeast8, 44 genes were detected exclusively by Yeast8, and 40 genes were detected exclusively by ecYeast8 (Fig. 3*C*). The genes exclusive to ecYeast8 were heme A synthase (*COX15*), nucleoside triphosphate pyrophosphohydrolase (*HAM1*), pentose phosphate pathway (*TKL1*, *RPE1*), alcohol dehydrogenase (*ADH1*), glucose uptake (*YRL446W*, *HXK1*, *HXK2*, *GLK1*, *EMI2*), isoprenoids and sterol biosynthesis

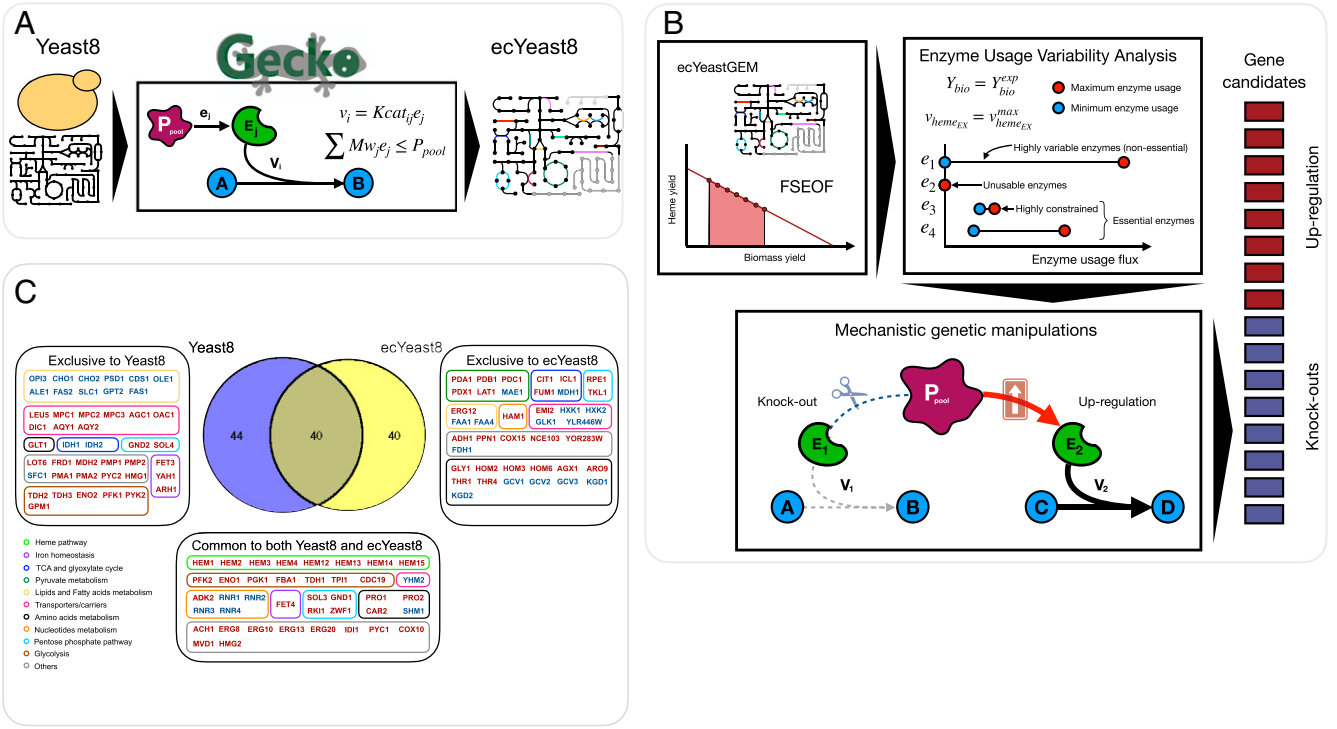


Fig. 3. The ecYeast8 model was used to find new targets for improved heme production. (*A*) Following use of the Yeast8, simulation using the enzyme-constrained model ecYeast8 was performed for increased heme production. (*B*) Following the adapted FSEOF approach (19, 22, 23), the enzyme usage variability analysis and mechanistic genetic manipulations for the individual gene modifications were used to refine the heme target list. (*C*) In simulations, the Yeast8 model identified 84 targets, and the ecYeast8 model identified 80 targets. Of the gene targets identified by the two models, 40 genes overlapped between Yeast8 and ecYeast8; 44 genes were identified by only the Yeast8 model, and 40 genes were identified by only the ecYeast8 model.

(*ERG12*), pyruvate metabolism (*PDA1*, *PDB1*, *PDC1*, *PDX1*, *LAT1*, *MAE1*), TCA cycle (*CIT1*, *MDH1*, *FUM1*), glyoxylate cycle (*ICL1*), glycine biosynthesis (*GLY1*, *AGX1*), glycine cleavage system (*GCV1*, *GCV2*, *GCV3*), fatty acids synthesis (*FAA1*, *FAA4*), L-lysine degradation (*KGD1*, *KGD2*), L-threonine metabolism (*HOM2*, *HOM3*, *HOM6*, *THR1*, *THR4*), phosphatase (*YOR283W*), polyphosphate metabolism (*PPN1*), formate dehydrogenase (*FDH1*), carbonic anhydrase (*NCE103*), and aromatic amino acids synthesis (*ARO9*) (Fig. 3*C*). Interestingly, among genes common to both Yeast8 and ecYeast8, *PRO1* was predicted to be downregulated by Yeast8, whereas it was predicted to be overexpressed by the ecYeast8 simulations. Experimental validation showed that deletion of this gene reduced the heme production drastically (Fig. 2*A*).

The small portion of positive Yeast8 genes (including *OPI3*, *CHO2*, *SLC1*, *PMA2*, *MPC3*, *MDH2*, *GLT1*, *FRD1*, *AQY1*, *AQY2*, *ALE1*, *SFC1*, and *AGC1*) were not detected by ecYeast8. However, these genes proved to improve heme production by the engineering genes one at a time (Fig. 2). These data can also be used for further improvement of the ecYeast8 model predictions.

Predicting Compatible Gene Combinations for Improved Heme Production. The list of genetic targets in Dataset S2 represents individual strategies for enhancing heme production, and we next used ecYeast8 to assess the viability of combining these strategies *in silico*. First, metabolic function redundancy was assessed by identification of identical genes in a genes-metabolites network (i.e., a bipartite graph that connects a metabolite with a gene if they are both involved in the same reaction). This allowed classification of gene targets in groups, where each gene group contains genes that are linked to the same metabolites according to the model. This grouping allowed a further reduction of the total number of candidates by discarding all genes that did not fit into any of the following categories: 1) gene target candidates with a unique metabolic function; 2) genes encoding for enzymes with the highest specific activity in a given group of redundant candidates for overexpression, due to their lower impact on the total protein burden for the cell; and 3) all gene candidates for deletion whose enzymes did not carry any flux in a reference flux distribution for optimized heme production (*SI Appendix*). Overall, this filtering procedure reduced the number of candidates from 80 to 71 genes.

We ranked the remaining modification targets according to the categories described above. Within each category, the fold-change in heme production was predicted for each individual target. An optimal mutant strain was then constructed *in silico* by implementing the remaining modifications in a sequential and cumulative way. Gene modifications that decreased the optimal production yield when compared to the previous iteration were discarded. This allowed us to obtain a list of "compatible" 58 gene target modifications that, according to the ecYeast8 model simulations, should yield a viable strain with enhanced heme production capabilities if they are combined (Dataset S3).

Constructing Compatible Gene Combinations for Improved Heme Production. We used the CRISPR-Cas9 toolbox developed for *S. cerevisiae* (37) to combine positive gene targets, which were predicted by the ecYeast8 model to yield higher production of intracellular heme, resulting in a viable strain. From our list of 58 compatible genes (Dataset S3), we overexpressed the *HEM13* gene first, as this gene had the maximum experimental effect (Fig. 2). The choice of the sequential targets to be combined with *HEM13* gene was evaluated experimentally. If the individual gene modification did not increase the output,

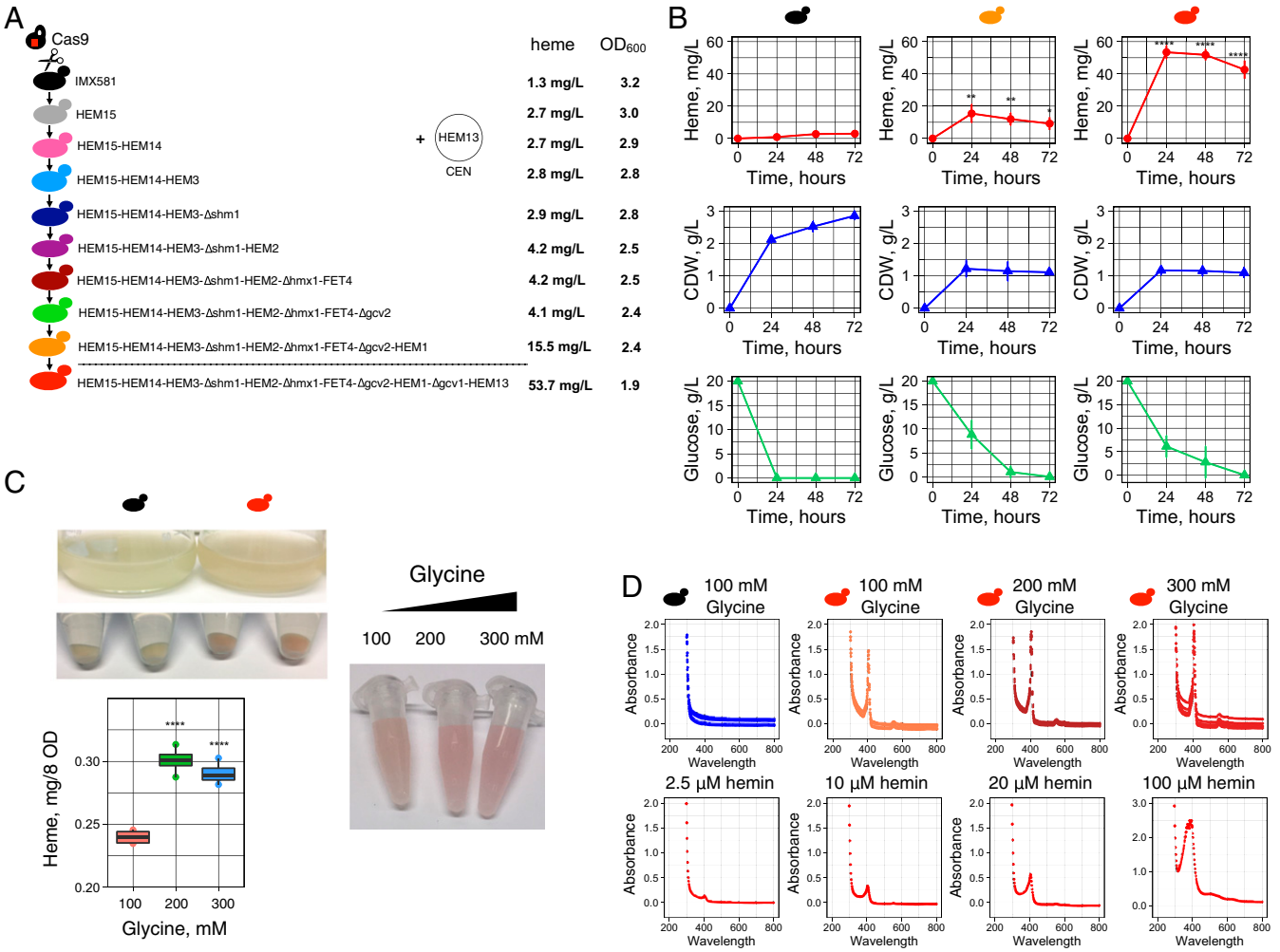


Fig. 4. CRISPR-Cas9 genome engineering for increased heme production. (*A*) The IMX581 strain carrying CRISPR-Cas9 gene integrated in the genome was used to carry the combinatorial engineering of heme gene targets deduced by Yeast8 and ecYeast8 genome-scale model. The gene integrations and deletions were performed using the gRNA constructs targeting different genome loci. The gene *HEM13* was overexpressed from the centromeric plasmid. The *HEM13* expression cassette was integrated into the genome in the final strain. Absolute heme (mg/L) was extracted from the entire biomass of the strains. (*B*) Heme production, CDW, and glucose consumption in different strains at 24, 48, and 72 h of cultivation in buffered SD ura- or SD with 2% glucose, 100 mM glycine supplemented with 0.1 mM Fe^{3+}. Four biological replicates (transformants) were used in the experiment. Error bars represent the SD. Commercial hemin was used to calibrate data. Strains: IMX581 carrying an empty vector; IMX581/*HEM15 HEM14 HEM3 Δshm1 HEM2 Δhmx1 FET4 Δgcv2 HEM1* carrying the *HEM13* centromeric plasmid; IMX581/*HEM15 HEM14 HEM3 Δshm1 HEM2 Δhmx1 FET4 Δgcv2 HEM1 Δgcv1* carrying *HEM13* expression cassette integrated into genome. Statistical analysis was performed using one-way ANOVA (*$P \leq 0.02741$, **$P \leq 0.00594$, ****$P \leq 0$). (*C*) The culture, cells, and cell extracts (obtained with oxalic acid treatment) of engineered strain IMX581/*HEM15 HEM14 HEM3 Δshm1 HEM2 Δhmx1 FET4 Δgcv2 HEM1 Δgcv1 HEM13* had a red color. Increasing the glycine amount from 100 to 300 mM resulted in a further increase in heme production. Statistical analysis was performed using one-way ANOVA (****$P \leq 0.00007$). (*D*) Spectral analysis of yeast extracts (obtained with oxalic acid treatment) shows the presence of the Soret peak (at 400 nm) characteristic to heme in IMX581/*HEM15 HEM14 HEM3 Δshm1 HEM2 Δhmx1 FET4 Δgcv2 HEM1 Δgcv1 HEM13* strain. Hemin (2.5, 10, 20, and 100 μM) spectra were used in comparison.

then it was declared a failure at that time (but kept as a possible modification for later experimentation with a new gene combination). Plasmids expressing single guide RNAs (sgRNAs) targeting different genomic loci were constructed (using the pMEL10 plasmid vector as a base; ref. 37) and used to integrate the expression cassettes of gene targets (as described in *SI Appendix*). The sequential gene modifications, which lead to sequential increases in heme production, are presented in Fig. 4*A*. The *HEM13* gene was expressed from centromeric plasmid under promoter *TEF1* of *S. cerevisiae* (Fig. 4*A* and *SI Appendix*, Fig. S1). The effects of introduced strain modifications were verified both by heme production measurement and by using Heme-LBB (Fig. 5 and *SI Appendix*, Figs. S9–S12). The Heme-LBB is a fusion protein of green fluorescent protein (GFP) and hemoglobin alpha-gamma subunits (*SI Appendix*) and was expressed under the copper-inducible promoter *CUP1* of *S. cerevisiae*. The biosensor fluorescence was designed to reflect the intracellular heme amount. Hemoglobin is a HCP, and heme incorporation during its translation determines correct folding (5, 6). Thus, we inferred that greater intracellular heme is associated with an increase in correctly folded GFP-hemoglobin protein that can be measured by the biosensor's activity (Fig. 5*A*). As the Heme-LBB is a new biosensor, it was used in parallel with direct heme measurement to study its response.

In the CEN.PK.113–11c strain background, the overexpression of the *HEM13* gene resulted in an average ~threefold increase in heme production (Fig. 2*C*). In contrast, the *HEM13* overexpression in IMX581 strain resulted in only 1.5-fold-higher heme production (*SI Appendix*, Fig. S1). Next, we integrated five heme biosynthetic genes (*HEM15*, *HEM14*, *HEM3*, *HEM2*, *HEM1*) into different genome loci step by step using CRISPR-Cas9, and this resulted in increased heme production (*SI Appendix*, Figs. S1 and S2). Our initial test using Heme-LBB with 5-ALA (which is the product of Hem1) in the medium

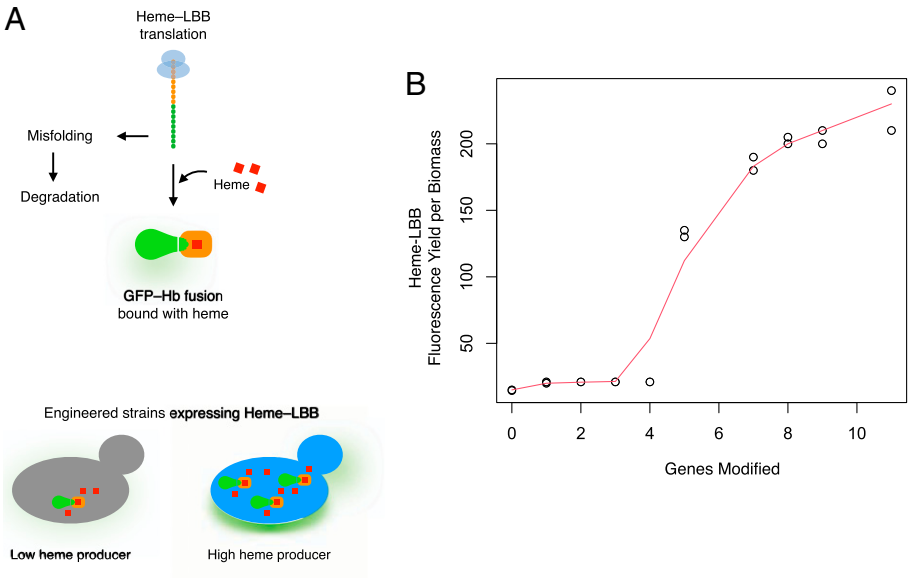


Fig. 5. Heme biosensor in engineered strains. (*A*) Heme-LBB is a fusion construct of GFP (highlighted in green) and hemoglobin (Hb, highlighted in orange). Heme (highlighted in red) is cotranslationally incorporated into the hemoglobin part of the biosensor polypeptide and promotes its correct folding. Heme-less biosensor molecules are misfolded and subjected to degradation. GFP-Hb fusion bound with heme is active and fluorescent. An increase in heme supply by the strain engineering will subsequently increase the number of correctly folded Heme-LBB molecules and, therefore, increase the strain's fluorescence. (*B*) Yield of Heme-LBB fluorescence per biomass with sequential heme-modeling targets engineered. Genes modified: 1: *HEM15*; 2: *HEM15*, *HEM14*; 3: *HEM15*, *HEM14*, *HEM3*; 4: *HEM15*, *HEM14*, *HEM3*, *Δshm1*; 5: *HEM15*, *HEM14*, *HEM3*, *Δshm1*, *HEM2*; 7: *HEM15*, *HEM14*, *HEM3*, *Δshm1*, *HEM2*, *Δhmx1*, *FET4*; 8: IMX581, *HEM15*, *HEM14*, *HEM3*, *Δshm1*, *HEM2*, *Δhmx1*, *FET4*, *Δgcv2*; 9: *HEM15*, *HEM14*, *HEM3*, *Δshm1*, *HEM2*, *Δhmx1*, *FET4*, *Δgcv2*, *HEM1*; 11: *HEM15*, *HEM14*, *HEM3*, *Δshm1*, *HEM2*, *Δhmx1*, *FET4*, *Δgcv2*, *HEM1*, *Δgcv1*, *HEM13*. Quantile regression with nondecreasing shape constraint (49) was used to estimate the biosensor response. To calculate the yield, the fluorescence of the Heme-LBB and the growth of each strain was monitored using a BioLector.

showed a response in the IMX581 strain carrying the *HEM15* gene expression cassette but no response in the control strain (*SI Appendix*, Fig. S14*A*). On the other hand, the biosensor activity increased with the engineered model targets that increase heme production (*SI Appendix*, Figs. S9, S11, S12, and S14*B*). The deletion of the *SHM1* gene combined with overexpression of heme genes (*HEM15*, *HEM14*, *HEM3*) resulted in a strain producing ∼fivefold more heme than the IMX581 control (*SI Appendix*, Fig. S2). The overexpression of *ACH1* did not result in improvement of heme production (*SI Appendix*, Fig. S3). Additional deletion of *FAA4*, *FDH1*, and *YLR446w* resulted only in a small improvement of heme production (*SI Appendix*, Figs. S4 and S5). The deletion of *GCV2* improved heme production in combination with only some genes (*SI Appendix*, Figs. S2, S4, and S7). The gene encoding the heme oxygenase (*HMX1*), which is responsible for heme degradation (38), was the integration locus we used for expression cassettes of the *FET4*, *ADH1*, and *ARH1* genes. The overexpression of *FET4* and deletion of *HMX1* was a better combination for heme improvement than overexpression of either *ADH1* or *ARH1* and deletion of *HMX1* (*SI Appendix*, Figs. S6 and S11). Further deletions of the *GCV2* and *GCV1* genes and integration of the *HEM1* and *HEM13* genes substantially improved heme production, resulting in the strain turning red (Fig. 4*C*) and the highest GFP fluorescence of the heme biosensor (Fig. 5 and *SI Appendix*, Fig. S12). Further evaluation of this production strain (IMX581 *HEM15 HEM14 HEM3 Δshm1 HEM2 Δhmx1 FET4 Δgcv2 HEM1 Δgcv1 HEM13*) using direct heme extraction and fluorescence measurement showed that it produced 53.5 mg heme per liter of the culture at 24 h of cultivation, which was 35.6 times higher than that of the initial strain, IMX581 (Fig. 4*B*). When normalized by the cell dry weight (CDW) (Fig. 4*B*), the constructed strain produced 70-fold more heme when compared to the initial strain. When the heme was extracted from the same amount of biomass, the production strain (IMX581 *HEM15 HEM14 HEM3 Δshm1 HEM2 Δhmx1 FET4 Δgcv2 HEM1 Δgcv1 HEM13*) contained 74.4 times more intracellular heme at 72 h of cultivation compared to the control strain, IMX581 (*SI Appendix*, Fig. S15). The fluorescence of the biosensor protein in the constructed strain was also the highest and was ∼20-fold higher than that of initial strain IMX581 (Fig. 5 and *SI Appendix*, Fig. S12). The best-performing strain also accumulated less biomass and consumed less glucose (Fig. 4*B*). Its growth rate was reduced by 40% (Fig. 4*B* and *SI Appendix*, Fig. S13), and its heme titer was 35-fold greater (Fig. 4*B*).

To test the possibility of a further increase in heme production, we studied heme produced in the IMX581 strain with genotype *HEM15 HEM14 HEM3 Δshm1 HEM2 Δhmx1 FET4 Δgcv2 HEM1 Δgcv1 HEM13* when cultured with elevated amounts of the glycine, the substate of Hem1 (Fig. 4*C*). As shown in Fig. 4*C*, the cell extracts of cultures grown on the medium supplemented with 200 or 300 mM glycine had a 25 or 20% greater heme, respectively. This was also accompanied with a darker red color of the yeast extracts (Fig. 4*C*). While both media with 200 or 300 mM glycine resulted in significantly higher heme production than the medium with 100 mM glycine (ANOVA, $^{****}P \leq 0.00007$), the difference in heme production between cultures grown in media with 200 or 300 mM glycine was not significant (ANOVA, $P \leq 0.19484$) (Fig. 4*C*). Unlike the control strain, the extracts of the production strain displayed a characteristic of heme Soret peak (at 400 nm) similarly to hemin, which was used as standard (Fig. 4*D*). Future studies should optimize heme production using glycine in fed-batch bioreactors and introduce the remaining gene modifications deduced by the model to improve heme production further.

Discussion

Heme is a cofactor of heme proteins and enzymes crucial for aerobic cell physiology (1). Free heme, heme proteins, and heme enzymes have been used in emerging technologies, such as flavoring

agents for artificial meat (39), blood substitutes (40), lithium-air batteries (10), and recently discovered chemical reactions (12, 13). High levels of intracellular heme are toxic to cells. Cytosolic heme is between 20 and 40 nM. Mitochondria tolerate higher concentrations of heme, roughly 30 μM (30). Higher levels of heme increase the production of hemoglobin and of P450 enzymes, which is apparently because the heme group insertion is essential for the proper folding and conformational stability of heme proteins (5, 6, 19, 41, 42). On the other hand, the overexpression of HCPs depletes the cellular heme pool and stresses the cell (18).

Recent advances in the GEM of *S. cerevisiae* (27–29) facilitate genome-scale identification of metabolic fluxes active in heme production, which can then be optimized to increase heme production. Linear and quadratic programming are needed for computing optimal, basic, feasible solutions and optimal interior solutions. Solving such problems is usually beyond the scope of humans.

Like *E. coli*, the yeast *S. cerevisiae* has GEMs of proven research supported by international communities of researchers (29, 43–45). In our study, to increase the cellular heme pool in the yeast *S. cerevisiae*, we used a metabolic modeling approach on the genome scale to maximize heme production by genetic modifications with *in silico* predictions and *in vivo* confirmation. Using FBA with Yeast8 and then enzymatically constrained ecYeast8 models (27, 29), we identified 84 gene as candidates to increase heme production. Our modeling suggested overexpressing 62 genes, downregulating 14 genes, and deleting 8 genes.

In the experimental phase of the study, several strategies were used. The strong constitutive promoter *TEF1* was used for overexpression of candidate genes. For the deletion and downregulation of genes, we used mutants from a collection of YKOs. Our one-gene-at-a-time experiments increased heme production by many interventions: strengthening glycolysis; improving the transport of pyruvate into mitochondria; improving the flow of acetyl-CoA into TCA cycle; overexpressing genes of the TCA cycle; modifying glycine-serine metabolism; and improving the transport of iron, water, and amino acids.

Then, additional modeling was performed to optimize combinations of gene modifications. Building on previous approaches to the prediction of gene overexpression targets (32, 46, 47), our study developed a procedure to identify some gene combinations that are compatible with the specified set of growth rates. With optimization approaches, increasing the flux for a reaction (or fluxes for reactions) allows other fluxes to change unless additional constraints are introduced to fix their values. FBA approaches have had difficulty accounting for, for example, protein burden, potential inhibitions by reaction products, and regulatory feedback loops.

In the first round of simulations with Yeast8, we identified genes that individually influence heme production. In the second round with ecYeast8 (with enzyme constraints), we developed *in silico* a viable mutant strain with improved heme production that had accumulated many positive modifications, successively added after having increased heme production above the previous maximum. In our model, we blocked the enzyme usage reactions for deletion targets, and for the overexpression targets, we doubled the enzyme usage. In laboratory experiments, the identified target combination was then engineered using CRISPR-Cas9. Our constraint-based model and our algorithm produced a list of 58 compatible genes. Thereafter, our implementation of changes sequentially chose the largest increase predicted by the model. If the individual gene modification did not increase the output, then it was declared a failure at that time (but kept as a possible modification for later experimentation). The first deletion of the *GCV2* gene did not increase output; at a later stage (after having introduced successful modifications), *GCV2* reappeared as a gene with maximum predicted increase, and it was (per our method) reintroduced, this time successfully. Increased heme production was positively and strongly associated with an increased activity of the newly developed Heme-LBB, as expected; the predicted increase in heme availability improves the cotranslational incorporation into hemoglobin. The Heme-LBB response to increased heme productivity was found to be dose dependent and sigmoidal, which is typical for biosensors. The developed biosensor provided the opportunity to measure heme *in vivo* without the need to extract heme for measurements. The biosensor activity in the constructed strains also assessed the expression of heme protein, which is useful for future work on the production of heme proteins using these strains.

With linear programming algorithms, our approach generated very interesting findings, which were not noticed previously in the literature. For example, our model found that heme biosynthesis is tightly coupled to central carbon metabolism with 80 genes, whose expression affects the heme production. Also, unexpectedly, the model implied that improved heme production could be achieved by reducing the lipid and deoxyribonucleotide triphosphates (dNTPs) biosynthesis and by increasing the activity of pentose phosphate pathway.

Our enzyme-constrained GEMs enabled us to develop a yeast strain with 70-fold more intracellular heme compared to the control strain when normalized per biomass. Our engineered strain produced 53.5 mg/L heme. Zhao et al. (17) achieved the intracellular production of 51.5 mg/L in *E. coli*, which is comparable with our yeast strain producing 53.5 mg/L total heme. Improving heme output by an order of magnitude in our strain required the simultaneous modification of 11 genes, which were selected through GEM simulation and laboratory experimentation. Our strain overexpressed the heme-biosynthetic genes *HEM15*, *HEM14*, *HEM3*, *HEM2*, *HEM1*, and *HEM13*, and it also overexpressed the low-affinity Fe (II) transporter of the plasma-membrane gene (*FET4*). In addition, we deleted the mitochondrial serine hydroxymethyltransferase gene (*SHM1*), the heme oxygenase gene (*HMX1*), and the two genes encoding subunits of the mitochondrial glycine decarboxylase complex (*GCV1* and *GCV2*). The constructed strain with 11 genetic modifications can be further engineered with the 58 genetic modifications predicted to be beneficial. However, the introduction of numerous genetic modifications in one strain risks off-target mutations, and the 11 implemented modifications already increased heme production by 70 times.

Materials and Methods

All the materials and methods are detailed in *SI Appendix*. These include preliminary target selection using Yeast8; reference flux distribution using ecYeast8; gene target selection using ecYeast8; identification of an optimal combination of targets using ecYeast8; media and growth conditions; genome engineering; determination of glucose concentration; CDW analysis; determination of heme concentration; and heme biosensor. Briefly, the Yeast8 metabolic model of *S. cerevisiae* was used to identify preliminary gene targets using FBA. Then, the ecYeast8 allowed the incorporation of enzyme constraints and informed the selection of gene targets. Intracellular heme was extracted with oxalic acid (18). For the cell dry weight (CDW), cells were collected on 0.45 μm cellulose-acetate filter paper (Satorius Biolabs). We developed the Heme-HBB as a synthetic fusion protein (consisting of α-globin, γ-globin, and GFP) to detect heme *in vivo*, validating its increasing response experimentally. The Heme-HBB construct was expressed under the control of the copper-inducible promoter *CUP1*.

Statistical Analysis. The statistical programs R (48) and Minitab 18.1 were used to analyze the data. The biosensor response was studied with quantile regression with a nondecreasing shape constraint (49, 50).

Data Availability. All the necessary scripts for model prediction and analysis used in this study have been deposited to GitHub and are available at https://github.com/SysBioChalmers/heme_production_ecYeastGEM/releases/tag/v1.0 (51) or through Zenodo at https://doi.org/10.5281/zenodo.6792435 (52).

ACKNOWLEDGMENTS. O.P.I. and D.P. were funded by a grant from Swedish Foundation for Strategic Research (RBP14-0055). I.D. has received funding from the European Union's Horizon 2020 Framework Programme for Research and Innovation Grant Agreement 720824. I.D., B.J.S., and J.N. were funded by the EU project DD-DeCaF (grant 686070) and the Novo Nordisk Foundation (grant NNF10CC1016517). Authors thank Professor Thomas Nyström for the BY4741 deletion strains, Honzhong Lu and Mario Beck for providing insight on the predictive method implemented with ecYeast8, Eduard Kerkhoven and Feiran Li for the numerous discussions on ecYeast8, and Jim E. Blevins and Louis H. Scott for scientific editing.

Author affiliations: [a]Department of Biology and Biological Engineering, Systems and Synthetic Biology, Chalmers University of Technology, Gothenburg, Sweden; [b]Novo Nordisk Foundation Center for Biosustainability, Chalmers University of Technology, Gothenburg, Sweden; [c]Department of Biotechnology and Biomedicine, Section for Synthetic Biology, Technical University of Denmark, Kgs. Lyngby, Denmark; [d]Novo Nordisk Foundation Center for Biosustainability, Technical University of Denmark, Kgs. Lyngby, Denmark; and [e]BioInnovation Institute, Copenhagen, Denmark

1. R. Bernhardt, Cytochromes P450 as versatile biocatalysts. *J. Biotechnol.* **124**, 128–145 (2006).
2. K. T. Patton, *Anatomy & Physiology* (Elsevier - Health Sciences, 2015).
3. I. U. Heinemann, M. Jahn, D. Jahn, The biochemistry of heme biosynthesis. *Arch. Biochem. Biophys.* **474**, 238–251 (2008).
4. L. He *et al.*, Antioxidants maintain cellular redox homeostasis by elimination of reactive oxygen species. *Cell. Physiol. Biochem.* **44**, 532–553 (2017).
5. A. A. Komar, A. Kommer, I. A. Krasheninnikov, A. S. Spirin, Cotranslational heme binding to nascent globin chains. *FEBS Lett.* **326**, 261–263 (1993).
6. A. A. Komar, A. Kommer, I. A. Krasheninnikov, A. S. Spirin, Cotranslational folding of globin. *J. Biol. Chem.* **272**, 10646–10651 (1997).
7. A. F. Palmer, M. Intaglietta, Blood substitutes. *Annu. Rev. Biomed. Eng.* **16**, 77–101 (2014).
8. S. Moradi, A. Jahanian-Najafabadi, M. H. Roudkenar, Artificial blood substitutes: First steps on the long route to clinical utility. *Clin. Med. Insights Blood Disord.* **9**, 33–41 (2016).
9. R. Z. Fraser, M. Shitut, P. Agrawal, O. Mendes, S. Klapholz, Safety evaluation of soy leghemoglobin protein preparation derived from *Pichia pastoris*, intended for use as a flavor catalyst in plant-based meat. *Int. J. Toxicol.* **37**, 241–262 (2018).
10. W. H. Ryu *et al.*, Heme biomolecule as redox mediator and oxygen shuttle for efficient charging of lithium-oxygen batteries. *Nat. Commun.* **7**, 12925 (2016).
11. E. N. Mirts, I. D. Petrik, P. Hosseinzadeh, M. J. Nilges, Y. Lu, A designed heme-[4Fe-4S] metalloenzyme catalyzes sulfite reduction like the native enzyme. *Science* **361**, 1098–1101 (2018).
12. S. B. Kan, R. D. Lewis, K. Chen, F. H. Arnold, Directed evolution of cytochrome c for carbon-silicon bond formation: Bringing silicon to life. *Science* **354**, 1048–1051 (2016).
13. F. H. Arnold, Directed evolution: Bringing new chemistry to life. *Angew. Chem. Int. Ed. Engl.* **57**, 4143–4148 (2018).
14. K. Sasaki, M. Watanabe, T. Tanaka, T. Tanaka, Biosynthesis, biotechnological production and applications of 5-aminolevulinic acid. *Appl. Microbiol. Biotechnol.* **58**, 23–29 (2002).
15. S. J. Kwon, A. L. de Boer, R. Petri, C. Schmidt-Dannert, High-level production of porphyrins in metabolically engineered *Escherichia coli*: Systematic extension of a pathway assembled from overexpressed genes involved in heme biosynthesis. *Appl. Environ. Microbiol.* **69**, 4875–4883 (2003).
16. S. Pranawidjaja, S. I. Choi, B. W. Lay, P. Kim, Analysis of heme biosynthetic pathways in a recombinant *Escherichia coli*. *J. Microbiol. Biotechnol.* **25**, 880–886 (2015).
17. X. R. Zhao, K. R. Choi, S. Y. Lee, Metabolic engineering of *Escherichia coli* for secretory production of free haem. *Nat. Catal.* **1**, 720–728 (2018).
18. J. K. Michener, J. Nielsen, C. D. Smolke, Identification and treatment of heme depletion attributed to overexpression of a lineage of evolved P450 monooxygenases. *Proc. Natl. Acad. Sci. U.S.A.* **109**, 19504–19509 (2012).
19. L. Liu, J. L. Martínez, Z. Liu, D. Petranovic, J. Nielsen, Balanced globin protein expression and heme biosynthesis improve production of human hemoglobin in *Saccharomyces cerevisiae*. *Metab. Eng.* **21**, 9–16 (2014).
20. M. Hoffman, M. Góra, J. Rytka, Identification of rate-limiting steps in yeast heme biosynthesis. *Biochem. Biophys. Res. Commun.* **310**, 1247–1253 (2003).
21. O. P. Ishchuk *et al.*, Improved production of human hemoglobin in yeast by engineering hemoglobin degradation. *Metab. Eng.* **66**, 259–267 (2021).
22. J. L. Martínez, L. Liu, D. Petranovic, J. Nielsen, Engineering the oxygen sensing regulation results in an enhanced recombinant human hemoglobin production by *Saccharomyces cerevisiae*. *Biotechnol. Bioeng.* **112**, 181–188 (2015).
23. K. Y. Hara *et al.*, 5-Aminolevulinic acid fermentation using engineered *Saccharomyces cerevisiae*. *Microb. Cell Fact.* **18**, 194 (2019).
24. H. Lopes, I. Rocha, Genome-scale modeling of yeast: Chronology, applications and critical perspectives. *FEMS Yeast Res.* **17**, fox050 (2017).
25. E. J. Kerkhoven, P. J. Lahtvee, J. Nielsen, Applications of computational modeling in metabolic engineering of yeast. *FEMS Yeast Res.* **15**, 1–13 (2015).
26. F. Zhang, J. Keasling, Biosensors and their applications in microbial metabolic engineering. *Trends Microbiol.* **19**, 323–329 (2011).
27. R. Ferreira *et al.*, Model-assisted fine-tuning of central carbon metabolism in yeast through dCas9-based regulation. *ACS Synth. Biol.* **8**, 2457–2463 (2019).
28. B. J. Sánchez *et al.*, Improving the phenotype predictions of a yeast genome-scale metabolic model by incorporating enzymatic constraints. *Mol. Syst. Biol.* **13**, 935 (2017).
29. H. Lu *et al.*, A consensus *S. cerevisiae* metabolic model Yeast8 and its ecosystem for comprehensively probing cellular metabolism. *Nat. Commun.* **10**, 3586 (2019).
30. D. A. Hanna *et al.*, Heme dynamics and trafficking factors revealed by genetically encoded fluorescent heme sensors. *Proc. Natl. Acad. Sci. U.S.A.* **113**, 7539–7544 (2016).
31. J. Zhang *et al.*, Combining mechanistic and machine learning models for predictive engineering and optimization of tryptophan metabolism. *Nat. Commun.* **11**, 4880 (2020).
32. H. S. Choi, S. Y. Lee, T. Y. Kim, H. M. Woo, In silico identification of gene amplification targets for improvement of lycopene production. *Appl. Environ. Microbiol.* **76**, 3097–3105 (2010).
33. N. E. Lewis *et al.*, Omic data from evolved *E. coli* are consistent with computed optimal growth from genome-scale models. *Mol. Syst. Biol.* **6**, 390 (2010).
34. G. Giaever, C. Nislow, The yeast deletion collection: A decade of functional genomics. *Genetics* **197**, 451–465 (2014).
35. T. Keng, *HAP1* and *ROX1* form a regulatory pathway in the repression of *HEM13* transcription in *Saccharomyces cerevisiae*. *Mol. Cell. Biol.* **12**, 2616–2623 (1992).
36. D. Chiabrando, F. Vinchi, V. Fiorito, S. Mercurio, E. Tolosano, Heme in pathophysiology: A matter of scavenging, metabolism and trafficking across cell membranes. *Front. Pharmacol.* **5**, 61 (2014).
37. R. Mans *et al.*, CRISPR/Cas9: A molecular Swiss army knife for simultaneous introduction of multiple genetic modifications in *Saccharomyces cerevisiae*. *FEMS Yeast Res.* **15**, fov004 (2015).
38. O. Protchenko, C. C. Philpott, Regulation of intracellular heme levels by HMX1, a homologue of heme oxygenase, in *Saccharomyces cerevisiae*. *J. Biol. Chem.* **278**, 36582–36587 (2003).
39. R. Fraser, P. O'Reilly Brown, J. Karr, C. Holz-Schietiger, E. Cohn, "Methods and compositions for affecting the flavor and aroma profile of consumables." US Patent 9700067B2 (2017).
40. F. Khan, K. Singh, M. T. Friedman, Artificial blood: The history and current perspectives of blood substitutes. *Discoveries (Craiova)* **8**, e104 (2020).
41. C. J. Reedy, B. R. Gibney, Heme protein assemblies. *Chem. Rev.* **104**, 617–649 (2004).
42. H. F. Ji, L. Shen, R. Grandori, N. Müller, The effect of heme on the conformational stability of micromyoglobin. *FEBS J.* **275**, 89–96 (2008).
43. X. Fang, C. J. Lloyd, B. O. Palsson, Reconstructing organisms in silico: Genome-scale models and their emerging applications. *Nat. Rev. Microbiol.* **18**, 731–743 (2020).
44. J. D. Orth *et al.*, A comprehensive genome-scale reconstruction of *Escherichia coli* metabolism–2011. *Mol. Syst. Biol.* **7**, 535 (2011).
45. B. S. Mienda, A. Dräger, "Genome-scale metabolic modeling of *escherichia coli* and its chassis design for synthetic biology applications" in *Computational Methods in Synthetic Biology. Methods in Molecular Biology*, M. A. Marchisio, Ed. (Humana, New York, NY, 2021), **vol. 2189**.
46. S. Ranganathan, P. F. Suthers, C. D. Maranas, OptForce: An optimization procedure for identifying all genetic manipulations leading to targeted overproductions. *PLOS Comput. Biol.* **6**, e1000744 (2010).
47. A. Chowdhury, A. R. Zomorrodi, C. D. Maranas, k-OptForce: Integrating kinetics with flux balance analysis for strain design. *PLOS Comput. Biol.* **10**, e1003487 (2014).
48. R Core Team, *R: A Language and Environment for Statistical Computing* (R Foundation for Statistical Computing, Vienna, Austria, 2016). https://www.R-project.org/
49. R. Koenker, *Quantile Regression* (Cambridge University Press, 2005), pp. 1–366.
50. R. Koenker, quantreg: Quantile regression. R package version 5.86 (2021). https://CRAN.R-project.org/package=quantreg.
51. I. Domenzain, B. J. Sánchez, Heme Production ecYeastGEM. https://github.com/SysBioChalmers/heme_production_ecYeastGEM/releases/tag/v1.0. Deposited 6 March 2022.
52. I. Domenzain, B. J. Sánchez, SysBioChalmers/heme_production_ecYeastGEM: Source code for publication. https://doi.org/10.5281/zenodo.6792435. Deposited 3 July 2022.

Paper V:

Computational biology predicts metabolic engineering targets for increased production of 102 valuable chemicals in yeast

Domenzain, I., Lu, Y., Shi, J., Lu, H., & Nielsen, J.

Manuscript, BioRxiv, 2023

Computational biology predicts metabolic engineering targets for increased production of 102 valuable chemicals in yeast

Iván Domenzain[1,2(‡)], Yao Lu[3(‡)], Junling Shi[4], Hongzhong Lu[5(*)], Jens Nielsen[1,2,6(*)]

[1] Department of Biology and Biological Engineering, Chalmers University of Technology, SE41296 Gothenburg, Sweden

[2] Novo Nordisk Foundation Center for Biosustainability, Chalmers University of Technology, SE41296 Gothenburg, Sweden

[3] College of Enology, Northwest A&F University, 3 Taicheng Rd, Xianyang, Shaanxi, China

[4] Key Laboratory for Space Bioscience and Biotechnology, School of Life Sciences, Northwstern Polytechnical University, 127 Youyi West Road, Xi'an, Shaanxi Province 710072, China.

[5] State Key Laboratory of Microbial Metabolism, School of Life Sciences and Biotechnology, Shanghai Jiao Tong University, Shanghai, China.

[6] BioInnovation Institute, Ole Maaløes Vej 3, DK2200 Copenhagen, Denmark

[‡] **These authors contributed equally to this work.**

(*) Corresponding authors.

Correspondence: hongzhonglu@sjtu.edu.cn, nielsenj@chalmers.se

Abstract

Development of efficient cell factories that can compete with traditional chemical production processes is complex and generally driven by case-specific strategies, based on the product and microbial host of interest. Despite major advancements in the field of metabolic modelling in recent years, prediction of genetic modifications for increased production remains challenging. Here we present a computational pipeline that leverages the concept of protein limitations in metabolism for prediction of optimal combinations of gene engineering targets for enhanced chemical bioproduction. We used our pipeline for prediction of engineering targets for 102 different chemicals using *Saccharomyces cerevisiae* as a host. Furthermore, we identified sets of gene targets predicted for groups of multiple chemicals, suggesting the possibility of rational model-driven design of platform strains for diversified chemical production.

One sentence summary:

Novel strain design algorithm ecFactory on top of enzyme-constrained models provides unprecedented chances for rational strain design and development.

Introduction

The accelerated rise of metabolic engineering, the rewiring of cells metabolism for enhanced production of metabolites[1], and synthetic biology, the assemble of novel synthetic biological components and their integration into cells[2], has enabled the development of microbial strains with increased production capabilities of chemicals from renewable feedstocks. These engineered microbes, also known as microbial cell factories (MCF), have been generated for production of multiple specialized compounds, such as pharmaceuticals[3,4], biofuels[5,6], food additives[7,8] and platform chemicals[9]. Most of these cases have relied on use of the bacterium *Escherichia coli* and the yeast *Saccharomyces cerevisiae* as platform cell factories. Despite success in development of many processes, complete development of MCFs usually takes several years of research and costs USD50M, on average, in order to bring a proof-of-concept strain forward for commercial production[10].

As metabolism is a complex and highly interconnected network, the time and resource intensive process of MCF development can be alleviated by the use of genome-scale metabolic models (GEMs) together with computational algorithms, aiming to find non-intuitive gene engineering targets for enhanced production[11]. Several methods for MCF design have been developed in past years and used to drive metabolic engineering projects such as production of lycopene[12,13] and lactate[14] in *E. coli*, and drug precursors in *S. cerevisiae* cells[15]. However, the most widely used methods for MCF design (MOMA[16], FSEOF[12], optKnock[17] and optForce[18]) tend to predict extensive lists of gene target candidates, and modelers often find themselves in need of imposing custom criteria to delimit the number of candidate gene targets to be tested, in order to reduce the amount of experimental work. Additionally, state-of-the-art GEMs tend to overpredict metabolic capabilities of cells due to the lack of kinetic and regulatory information in their formulation, hindering their applicability for further quantitative evaluation and comparison of predicted metabolic engineering strategies. Kinetic models have also been used for the development of strain design algorithms, such as k-OptForce[19], however, the limited size of this kind of models impedes prediction of metabolic gene targets in a genome-scale[20].

Here we present a computational method (ecFactory) for prediction of optimal metabolic engineering strategies, that circumvents the problem of arbitrary selection of the number of gene candidates by leveraging the vast amount of enzymatic capacity data, together with the improved phenotype prediction capabilities, of enzyme-constrained metabolic models (ecModels, generated by the GECKO toolbox)[21]. The performance of ecFactory was systematically tested and evaluated by comparing the predictions with experimental data for multiple study cases. Using this method we identified gene targets for increased production of 102 different chemicals in *S. cerevisiae*, enabling identification of gene targets common to

multiple groups of products, suggesting the opportunity for development of platform strains that can be used for diverse chemical production. Moreover, our analysis quantitative estimation of enzyme- and substrate- limitations for production of the 102 studied chemical products. To enable wider utilization of these results by the community, we established a web-based resource for accessible query and visualization of the gene target predictions in the context of Metabolic Atlas, and we expect this resource to facilitate significant advancements in development of yeast MCFs through metabolic engineering.

Results and discussion

Modelling production of 102 chemical products in yeast

A list of 102 industrially relevant natural products, whose metabolic production pathways are known and reported in the literature, was collected. Products were grouped into 10 different families according to their chemical characteristics: amino acids (26), terpenes (22), organic acids (15), aromatic compounds (9), fatty acids and lipids (9), alcohols (8), alkaloids (6), flavonoids (5), bioamines (2) and stillbenoids (1). From these, 50 products were found to be native metabolites in *S. cerevisiae*, whilst 52 products were identified as heterologous, according to an enzyme-constrained metabolic model for yeast (ecYeastGEM v8.3.4)[21]. A summary of the chemical classification of products is shown in **Fig. 1A** and **supp. table S1**. Production pathways were reconstructed for all these heterologous products and incorporated into ecYeastGEM, taking energy and redox requirements as well as reported kinetic data into account (see **Materials and Methods**). All of the 53 reconstructed heterologous pathways are described in **supp. table S2**.

In silico assessment of production capabilities for 102 chemicals in yeast using metabolic modeling with enzyme constraints

The production capabilities of *S. cerevisiae* were quantitatively explored, using both YeastGEM and ecYeastGEM, by computing optimal production yields for all of the 102 studied chemicals, constrained by low and high glucose consumption regimes (1 mmol/gDw h; and 10 mmol/gDw h) and biomass production rates spanning the range between zero and a maximum attainable value, using flux balance analysis (FBA) simulations[22].

As FBA relies on optimality principles, usually assuming maximization of cellular growth as a cellular objective[23], there is a trade-off between biomass formation and accumulation or secretion of products of interest. Yeast has evolved the ability to switch to mixed respiro-fermentative metabolic regimes when

nutrients are available in excess, favoring enzymatic efficiency over biomass yield on substrate[24–27]. As ecModels account for a limited enzymatic machinery in cells, different production capabilities are predicted when changing from low to high glucose uptake rate, in contrast to classic GEMs, that solely rely on stoichiometric constraints. This additional constraint results in a different production phase-plane as illustrated by **Fig. 1B**, i.e., instead of the standard linear trade-off between product formation and biomass formation there will be a regime where the product formation is limited by the protein constraint. Furthermore, the phase-plane becomes dependent on the glucose-uptake rate, such that at high glucose consumption the ecModel predicts a protein-limited regime of production, yielding lower production levels and biomass formation per unit of glucose. Protein limitations may also arise at low glucose consumption levels, for cases in which the production pathways for the chemical of interest involve inefficient enzymes (low specific activity). This introduces enzymatically unfeasible regions in the production space of a cell, indicated by the grey region in **Fig. 1B**. A typically protein-constrained production landscape with a region of difference between YeastGEM and ecYeastGEM predictions in the low glucose regime is shown for the alkaloid choline in **Fig. 1C**. In contrast, a production landscape solely governed by stoichiometric constraints at low glucose levels is shown in **Fig. 1D** for the polyamine putrescine. Additional examples of yield plots for chemicals belonging to all studied families can be found in **Fig. S1**.

Highly protein-constrained products were found by identifying those chemicals whose maximum production level demands the totality of the available enzyme mass in the model, at low levels of glucose consumption. In total, 40 out of the 53 analyzed heterologous products were found to be highly protein-constrained, in comparison to production of native metabolites, for which just 5 products were classified as part of the same group (**Fig. S2A**). Furthermore, strong protein limitations arise often for groups of heterologous chemicals derived from a native pathway with high enzymatic demands, such as terpenes and flavonoids, derived from the mevalonate pathway. On the other hand, few strongly protein-limited products were found amongst families connected to native biosynthetic processes, such as amino acids, organic acids and diverse alcohols (**Fig. S2B**). Protein constrained models offer the possibility of computing optimal costs of chemical production both in terms of substrate and required protein mass. Minimal protein and substrate mass costs per unit mass of product were computed for each of the 102 products (see **Materials and Methods** for further details), as has been previously suggested by other computational work[28]. **Fig. 1E** shows that a positive correlation between these two production costs exists, allowing the identification of slightly and highly constrained groups of products, with an overrepresentation of native products (amino acids, organic acids and some alcohols) in the former group, and heterologous chemicals (terpenes, flavonoids and some aromatic compounds) in the latter. This plot shows that for heterologous products, it is usually necessary to invest on improving enzyme properties, i.e., increase their catalytic efficiency,

whereas for native products it is predominantly stoichiometric constraints that should be considered for minimizing costs. Moreover, it was found that slightly constrained products tend to be lighter, in terms of molecular weight, than those in the highly constrained group. This is also suggested by substrate costs, as larger organic molecules require more carbon to be formed, notwithstanding, this also suggests that a heavier enzymatic burden is needed for assembly of large molecules, as it is likely that additional, and less efficient, enzymatic steps are involved in their synthesis.

The effect of increasing enzyme catalytic efficiency for improving production levels was explored with FBA simulations with ecYeastGEM at different activity levels of rate-limiting enzymes. For highly protein-constrained products, such as the alkaloid psilocybin, a monotonic linear decrease of the substrate cost is observed when decreasing the total production protein cost by enhancing the activity of the heterologous tryptamine 4-monooxygenase (P0DPA7). **Fig. S3A** shows that when the P0DPA7 catalytic efficiency is increased by 100-fold, the total oxygen consumption is predicted to increase by 75%, which suggests that reducing the protein burden of the psilocybin biosynthetic pathway releases protein mass that can be used by the cell to meet its energy demands by an increased respiratory rate. Overall, this metabolic rewiring shifts the psilocybin production space in a direction of higher product yields (**Fig. S3B**). However, the product yield is still low indicating that other enzymes in the pathway may have to be improved to further increase yield. A similar behavior was obtained for the case of valencene, a moderately protein-constrained terpenoid, by increasing activity levels of the sole heterologous limiting enzyme, terpene synthase (S4SC87), from 1 to a 100-fold. A positive correlation was also observed between substrate and protein costs for this product (**Fig. S3C**), however, lower slopes in the production cost space were obtained for higher activity values of S4SC87. **Fig. S3D** illustrates that increased activity of this limiting enzyme reduces the enzymatically unfeasible region of the valencene production space, bringing its optimal production line closer to the stoichiometrically constrained limit (blue and dark red lines).

In sum, model predictions indicate that heavily protein-constrained biosynthetic pathways could result in the increase of protein and substrate costs of production. This kind of pathways require resources from the limited cellular enzymatic machinery, hence, the substrate-efficient respiratory pathway for energy production is compromised in favor of substrate-inefficient fermentative pathways, which reduces the protein burden necessary for sustaining cellular energy levels.

An integrative constraint-based method for prediction of metabolic engineering strategies

The flux scanning with enforced objective function algorithm (FSEOF)[12] has been extensively used for identification of metabolic engineering targets in yeast, due to its implicit consideration of the tradeoff between biomass and metabolite production. It is of particular interest to explore this method in the context of ecModels as variable energetic and biosynthetic requirements may induce a complete change of the cellular behavior. Therefore, engineering strategies that minimize the substrate and protein costs for optimal bio production can be predicted, furthermore, predictions have boosted heme accumulation in yeast cells by 70-fold[29]. In order to ensure predictive robustness and minimizing the number of false positives among predictions, we revised and systematized this approach and developed ***ecFactory***, a multi-step constrained-based method for prediction of engineering gene targets for enhanced biochemical production, based on the principles of FSEOF and on the ability of ecModels to compute enzyme demands for biochemical reaction, providing systematic criteria to predict an optimal minimal set of modifications for increasing production of target metabolites.

In summary, ecFactory consists of three basic steps: 1) prediction of gene expression scores, indicating intensity and directionality of genetic modifications; 2) discard gene targets encoding for unfavorable enzymes (redundant, low efficiency) and; 3) Obtention of a minimal combination of modifications required for driving cells from optimal biomass formation to a metabolic production regime. The overall objective of this method is to obtain a reduced list of gene targets, focusing on the optimal strategies for enhanced production by taking enzyme allocation and connectivity into account. All the constituent steps of the ecFactory method are illustrated in **Fig. 2** and explained in detail in the **Materials and Methods** section of the **Supplementary Materials**.

Furthermore, the classification of targets according to the characteristics of their respective enzymes (illustrated by **Fig. S4**), facilitates a deeper understanding of the predicted optimal metabolic engineering strategies. The list of 12 gene targets for 2-phenylethanol (**Table S4**) suggests that, in order to increase production of this chemical, enzymes that are optimal for providing the necessary metabolic precursors and cofactors are predicted as targets for overexpression. Knock-down and knock-out targets aim to direct the metabolic flux towards optimal production while reducing the formation of biomass precursors in excess (glycerolipids in this case).

Enzyme constraints enable identification of optimal combinations of genetic modifications for 102 chemicals in yeast

The ecFactory method was used to predict gene targets for enhanced production of each of the 102 chemicals. The method proved to be effective at returning predictions for all cases, while reducing the number of candidate gene targets at each of its sequential steps. The distributions of the predicted number of gene targets per product (shown in **Fig. 3A)** shows the major contribution of classifying targets according to their enzymatic characteristics (step 2) at reducing the number of predicted OEs, KDs and KOs. On average the first step of the method (FSEOF), running on ecYeastGEM, predicted 85 gene targets per product (28 OEs, 42 KDs and15 KOs), the number of targets is then reduced by the following steps by 73%, as only optimal candidates are returned by the ecFactory algorithm (7 OEs, 9 KDs and 5 KDs per product, on average). Notably, predictions reveal that increasing production of protein-limited and heterologous chemicals require significantly more genetic modifications, compared to substrate-limited and native products (p-values = $1.16x10^{-5}$ and $2.3x10^{-3}$, respectively, under a one-sided two-sample Kolmogorov-Smirnov test) as shown in **Fig. S5**. These differences are caused by the large number of gene knock-downs and knock-outs that are required to change the energy production strategy from cellular respiration to a fermentative metabolism, so that the limited cellular enzyme capacity can be optimally allocated to the final production reaction steps, which tend to be inefficient for these kinds of products. A more detailed presentation of results, by chemical family, method steps, and target types, is available in **supp. table S3**.

Overall, 150 endogenous genes in yeast are predicted as OE target for at least one of the modeled products; 88 different genes are predicted as KD targets and 129 as KO targets. More than 50% of the targets predicted for OE, KD and KO are specific to one or two of the 102 products (Fig.3B, Fig. S6A and Fig. S6C). Nonetheless, small sets of genes are predicted as targets for a high number of products (promiscuous targets), spanning almost all chemical classifications in this study. Genes encoding for reaction steps in the pentose-phosphate pathway and pyruvate metabolism, together with PFK2 in the glycolysis pathway, are predicted as the most common OE targets across products; the most common KD and KO gene targets encode for enzymes in the TCA cycle, oxidative phosphorylation and synthesis of biomass precursors (steroids, glycerolipids, nucleotides and amino acids), as shown in **Fig. 3C**, **Fig. S6B** and **Fig. S6D**, suggesting a global strategy of redirecting carbon flux into heterologous pathways and alternative energy production mechanisms.

In silico predictions capture successful metabolic engineering strategies in yeast

It was found that 7 out of the 12 predicted gene targets to increase 2-phenylethanol have been previously engineered in *S. cerevisiae*, *Yarrowia lipolytica and Kluyveromyces marxianus* strains with enhanced 2-phenylethanol production levels[30–32] (**Table S4**), indicating that ecFactory predictions can be capable of

capture targets proposed by rational engineering approaches. As another example of experimentally validated predictions, the case of spermine is of particular interest. In this case, the ecFactory method was able to capture 9 of the implemented targets (MAT, ODC, SPE2, SPDS, MEU1 APT2 and PRS for OE; FMS1 and CAR2 for KO) in a successfully engineered strain for spermidine production, an immediate precursor of spermine[33]. It was also found that the experimental implementation of a heterologous cytosolic ornithine cycle was resembled by a general predicted overexpression of its native mitochondrial version.

These particular results suggest that the method is able to capture the underlying logic of highly complex rational engineering approaches that require the coordination of multiple sectors of metabolism, as shown by **Fig. S7**. Overall, the predicted gene modifications aim to increase spermine biosynthesis by overexpression of the whole ornithine cycle, a direct precursor, together with the Yang cycle and some steps in the pentose phosphate pathway (PPP) in order to increase S-adenosyl-L-methionine, another important precursor of polyamines. Interestingly, when focusing on the final predictions for this product (targets in step 3), just 5 of the 8 aforementioned genes were classified as optimal targets for spermine production (SPDS, ARG8 andARG5,6 OEs, together with FMS1 and CAR2 KO). This suggests that, according to enzyme capacity and metabolic connectivity, it is possible to reduce complex rational metabolic engineering strategies, to fewer modifications on crucial reaction steps in pathways that need to be rewired and coordinated, one of the purposes for which this method was designed.

In order to validate the quality of the ecFactory predictions, we searched the literature for independent experimental studies in *S. cerevisiae* that have been successful at increasing production levels of chemicals included in our list. Gene modifications validated for diverse chemicals were found to be predicted as optimal gene targets by ecFactory, shown in **Table 1**. Interestingly, several of these targets are common to multiple products. In total, 28 predicted different gene targets were found as experimentally validated across 22 products, which are also part of different chemical classes. The most repeated genes among these targets correspond to overexpression in the ergosterol, mevalonate, shikimate and polyamine biosynthesis pathways.

Table 1.- Predicted gene targets with experimental validation.

Product	**Chemical class**	**Validated overexpressions**	**Validated KD/KOs**
amorphadiene[34]	terpene	HMG1,ERG8,ERG12,MVD1,ERG20	
artemisinic acid[3]	terpene	HMG1	

β-amyrin[35]	terpene	ERG9,ERG20,ERG20,ERG9,HMG1, ERG8,ERG13,ERG12,MVD1	
β-ionone[36]	terpene	BTS1	
Cinnamoyl-tropine[37]	alkaloid	SPE1	
cis,cis-muconate[38]	organic acid	SPE1	ZWF1
α-farnesene[39]	terpene	HMG1,ERG20	
geraniol[40]	terpene	HMG1	
glutathione[41]	amino acid	GSH1	
Hydroxy-mandelic acid[42]	aromatic	ARO1	
malate[43]	organic acid	MDH2	
mandelic acid[42]	aromatic	ARO1	
miltiradiene[44]	terpene	BTS1	
nootkatone[45]	terpene	ERG20,HMG1	
ornithine[46]	amino acid	GDH1	
2-phenylethanol[30–32]	alcohol	ARO2,PHA2,ARO10,ARO1,ARO4, ARO7,ZWF1	
pyruvate[47]	organic acid		PDC5,PDC6
2,3 R-R-butanediol[48]	alcohol	PDC1	
santalene[39]	terpene	HMG1,ERG20	
spermine[33]	bioamine	GDH1,SPE2,SPE3,MEU1	FMS1,CAR2
squalene[49]	terpene	HMG1,ERG20,ERG9,ERG8,ERG12, MVD1	
valencene[50]	terpene	HMG1,ERG8,ERG12,ERG20,MVD1	

These similarities at the gene and pathway level among predictions for different chemical products, suggest the existence of metabolic engineering strategies capable of providing the necessary precursors for increasing production of groups of chemicals. This kind of strategies have been sought in experimental metabolic engineering, following rational approaches, and have proved to be successful for the development of platform yeast strains for production of different groups of molecules such as opioids[4] and other alkaloids[51,52], polyketides[53] and terpenes[39,54]. Furthermore, cumulative combination of individual genetic modifications in a production strain is needed for achieving meaningful flux towards the desired chemical[29], therefore, it is desirable to identify multiple gene targets, encompassing multiple metabolic pathways, that constitute the chassis for robust and diversified chemical production.

Gene targets common to all products in a given chemical family were sought for all cases in this study. The only chemical family with common predicted targets was found to be flavonoids, with 9 KDs (ADO1, ATP19, IDP1, LPD1, MAE1, MDH2 MET6, PPA2 and SAH1) and 7 KOS (CAR2, FAA, FAA4, FDH1, RNR1, RNR3 and RNR4). This combination of targets reveals an engineering strategy that decreases the TCA cycle and respiratory fluxes, the amount of carbon going towards acetyl-CoA and posterior fatty acid synthesis, synthesis of amino acids derived from 2-oxoglutarate and nucleotides biosynthesis. Altogether, this shows an optimal way of allocating carbon flux and the limited enzymatic machinery of yeast for the biosynthetic pathways producing catechin, genistein, kaempferol, naringenin and quercetin. Nevertheless, the impact of these modifications on other biological processes, such as regulatory networks, is not accounted for in the metabolic model and should be further assessed.

Model-driven design of platform strains for diverse chemical production

As highly promiscuous gene targets, for all kind of modifications, were found to be predicted for products present in most of the studied chemical families, other sets of targets common to groups of multiple products may exist among the ecFactory predictions. In order to systematize the analysis of gene target profiles across products, the 102 lists of targets were represented as mathematical vectors (see **Materials and Methods** section of the **Supplementary Materials** and **Figure 4A** for further details). Highly similar gene expression vectors were identified using the t-distributed stochastic neighbor embedding method (t-SNE), which is suited for visualization and identification of clusters in high dimensional datasets[55]. Two-dimensional representation of t-SNE results facilitated identification of 8 different clusters of target vectors, representing different groups of products. Product clusters are shown in **Figure 4B**. Notably, gene targets common to all products in a group were found for all clusters (**Table 2**).

Table 2.- Shared gene targets within each cluster of products.

Cluster	Chemical Products	Shared KO targets	Shared KD targets	Shared OE targets
1	betaxanthin, caffeic acid, vanillin β-glucoside, β-ionone, glycyrrhetinic acid, miltiradiene, lycopene, taxadien-α-yl acetate, protopanaxadiol, genistein, quercetin, catechin, kaempferol, patchoulol, oleanolate, lupeol	RNR1, RNR4, RNR3, CAR2. FAA4, FAA1, FDH1	SAH1, ARG5,6, MET6, LPD1, ADO1, MAE1, ARG7, MDH2, ARG8, ATP19	NA

2	β-carotene, cinnamoyltropine, ARA, DHA, EPA, astaxanthin, psilocybin, docosanol	RNR1, RNR4, RNR3, CAR2. FAA4, FAA1, FDH1	IDP1, ARG5,6, LPD1, MAE1, MDH1, ARG7, PPA2, MDH2, ARG8, ATP19	NA
3	ergosterol, squalene, santalene, farnesene, amorphadiene, limonene, geraniol, artemisinic acid	NA	LPP1	PDB1, PDA1, PDX1, ERG12, ERG8, LAT1, MVD1
4	Itaconic acid, glutamine, proline, putrescine, spermine	NA	LPP1	PDB1, PDA1, PDX1, LAT1
5	valencene, nootkatone, linalool, β-amyrin	NA	ARG5,6, ARG8	ERG12, ERG8, MVD1
6	tryptophan, adipic acid, cis-muconate, hydroxymandelic acid	MAE1	LPP1	ARO4
7	phenylalanine, 2-phenylethanol, mandelic acid, cinnamate	MAE1	LPP1	ARO4, ARO1, ARO2, SOL3, GND1, ZWF1, PHA2, ARO7
8	Free-fatty acids, oleate, palmitoleate	NA	LPP1, ARG5,6, MAE1, CAR2, ARG8	CDC19, BPL1, SOL3, GND1, PDC1, ACS2, PPA2, ZWF1, ACC1, ALD6

In general, these clusters are composed by products that belong to different chemical families, with the exception of cluster 3 and 5, composed mostly by terpenes, and cluster 8, formed just by lipid compounds. Mapping product origin (native or heterologous) and protein limitations information into the clustering results showed that, clusters 1 and 2 are composed by heterologous and highly protein-constrained products belonging to different compound classes; terpenes whose production is constrained by substrate availability tend to group together, in cluster 3; and most native products, despite their protein limitations, do not fall into the identified clusters. Altogether, this shows that metabolic engineering strategies for the different product clusters are defined by gene modifications that are related with redirecting flux and energy from central metabolism to the final specific heterologous pathways. This suggests that shared molecular characteristics between products (i.e., chemical classification of products) might not be the most decisive aspect when designing genetic modification strategies for enhanced production of multiple chemicals (platform or chassis strains).

In order to understand the particular metabolic rewiring required by each platform strain designed with the aid of the cluster analysis, turnover rates were calculated for the 12 main precursor metabolites in central carbon metabolism (D-glucose-6-phopshate, D-fructose-6-phosphate, ribose-5-phosphate, erythrose-4-phosphate, glyceraldehyde-3-phosphate, 3-phosphoglycerate, phosphoenolpyruvate, pyruvate, acetyl-CoA, 2-oxoglutarate, succinyl-CoA and oxaloacetate)[11] using FBA simulations for different scenarios, optimal biomass formation and optimal production of each of the studied chemicals. Fold-changes were then computed for each of the precursor turnover rates, by comparing the optimal production flux distributions to their optimal biomass formation counterpart, for all 102 production scenarios. In this way fold-changes higher than one indicate that, for increased production, the overall flux towards a precursor should be upregulated, in comparison to a wild-type metabolic state, while fold-changes lower than one imply that the flux towards a precursor needs to be down-regulated (see **Materials and Methods** section in the **Supplementary Materials**).

Figure 4C shows that significant upregulation of flux towards erythrose-4-phosphate (E4P) and pyruvate, moderate upregulation of phosphoenolpyruvate (PEP), a drastic decrease in ribose-5-phosphate (R5P) and α-ketoglutarate (AKG) turnover rates and, a moderate down-regulation of the flux towards oxaloacetate (OXO), acetyl-CoA and succinyl-CoA should be combined to achieve optimal production levels of the products in clusters 1 and 2. Additionally, it can be seen that fluxes towards precursors located downstream from pyruvate (TCA cycle intermediates and acetyl-CoA) are needed to be downregulated for products in these clusters. This can be explained by a lower demand of building blocks, due to the decrease of biomass formation rate in a production scenario. Moreover, as all products in these clusters were found to be protein-limited, a predicted coordinated down-regulation of the lower section of central carbon metabolism suggests that forcing a fermentative regime, in which most of the energy is produced by glycolysis to minimize the protein burden induced by cellular respiration, thus, leaving room for expression of inefficient heterologous enzymes, offers the optimal conditions (metabolic mode) for production of these chemicals.

For the case of products in cluster 3, predictions indicate that a metabolic rewiring that induces significant upregulation of R5P, E4P and pyruvate production, and intense down-regulation of the flux towards and α-ketoglutarate is needed to improve production of these terpene compounds (**Figure 4D**), suggesting that an increased supply of NADPH (produced in the first steps of the pentose phosphate pathway, preceding ribose-5-phosphate) is needed for these products. The gene target profiles for the bioamines putrescine and spermine were found to cluster together with their precursor amino acids proline and glutamate, as well as itaconic acid (cluster 4). **Figure S7A** shows that genetic modifications common to all products in this cluster

cause only moderate changes in the turnover rate of central carbon metabolism precursors, mostly for those in lower glycolysis, indicating that the optimal production mode for these products does not differ significantly from a wild-type optimal growing metabolic strategy. A strong requirement for increased flux towards E4P was found to be common to all terpenes, despite the protein limitations involved in their production pathways, as shown by Figures **4C**, **4D** and **S7B**.

Production of native and heterologous products derived from the shikimate pathway, those in clusters 6 and 7, were found to require an increase of flux towards the immediate precursors E4P and PEP, together with enhanced NADPH supply, provided by an increased flux to R5P, and a reduction of the metabolic turnover of precursors located downstream of PEP, in order to maximize carbon conversion (**Figure S7C**). Finally, significant increase of acetyl-CoA turnover, together with a moderate upregulation of the pentose-phosphate pathway for increased NADPH flux, was found to be the optimal reprogramming strategy for production of free fatty acids, oleate and palmitoleate (**Figure S7D**), resembling previous successful work in yeast cells[56].

The set of common target predictions for a given cluster of products provides a modulated gene expression program capable of rewiring central carbon metabolism for increased production of key precursor metabolites. Implementation of these predictions in yeast cells can be used to drive the development of platform strains, specialized in providing the production scaffold for multiple chemicals. Platform strains can then be transformed into product-specific ones by introducing the necessary heterologous genetic components. This platform-based procedure will potentially reduce the resources and efforts involved in the development of next-generation cell factories.

Web-based resources for exploration of metabolic engineering targets in S. cerevisiae

Predicted gene targets for increased production of the chemicals in this study were incorporated into metabolic atlas for visualization in a metabolic network context. **Figure 5** shows the gene modifications for improved patchoulol production in the central carbon metabolism of yeast as an example, where genes indicated for OE, KD and KO can be found. Furthermore, metabolic maps for other pathways, even in secondary and intermediate metabolism, are also available. Visualization options for the 102 products can be found at: www.dev.metabolicatlass.org. Additionally, in order to facilitate the utilization of the ecFactory method, interactive tutorials for prediction of engineering targets for 2-phenylethanol and heme production in yeast are available as MATLAB live scripts at: https://github.com/SysBioChalmers/ecFactory/tree/main/tutorials.

Conclusions

Here we demonstrated that, by accounting for enzyme limitations, the use of metabolic models for quantitative prediction in metabolic engineering can be extended and improved. Enzyme-constrained models enabled assessment of the impact of enzyme capacity on the total protein and substrate costs of chemical production in cell factories, and reduction of the number of gene engineering targets for increased production predicted by stoichiometric constraint-based methods to a minimal optimal set of modifications. The model ecYeastGEM was used to predict gene engineering targets for enhanced production of 102 chemical products with yeast cells, including native and heterologous biochemicals with distinct chemical characteristics. Predictions showed to resemble complex engineering strategies that involve coordinated modulation and coordination of multiple pathways. Notably, supportive experimental evidence was found in the literature to verify the gene target predictions in 22 of the studied chemicals.

Sets of gene targets common across products were identified for 8 different groups of chemicals, inferred with a clustering algorithm. Flux balance analysis simulations indicate that, these core genetic modifications represent the expression tunning profiles, needed to rewire the central carbon metabolism of yeast towards increased production of the main metabolic precursors required by groups of valuable chemicals. By visualizing the 8 different rewiring schemes we learned that clustering of products according to their gene target predictions obeys to combinations of these three basic factors: 1) protein burden induced by the specific production pathways and its impact on energy production; 2) the metabolic precursor that provides the main carbon flux for final product formation; 3) products that require increased NADPH flux levels. Thus, the presented approach suggests the advantages of using of enzyme-constrained models for design and understanding of platform strains optimized for diverse chemical production. Nonetheless, expanding the scope and number of chemicals and host organisms for this kind of large-scale studies might help to unveil additional core principles for rationally engineering of metabolism.

We envision that the tools and methodology developed in this study will contribute to accelerate development of robust and efficient microbial strains both for specialized and also versatile production of valuable chemicals, promoting the conversion from petrol a bio-based economy.

Acknowledgments

This project has received funding from the Novo Nordisk Foundation (grant no. NNF10CC1016517), the Knut and Alice Wallenberg Foundation, and the European Union's Horizon 2020 research and innovation program with projects DD-DeCaF and CHASSY (grant agreements No 686070 and 720824). This project was also supported by The Shanghai Pujiang Program and Grants 22208211 from the National Natural Science Foundation of China (NSFC).

Author Contributions

I.D. wrote the draft manuscript, developed the software and methods and analyzed results. Y.L. performed the literature review, constructed the heterologous pathways in ecYeastGEM and contributed to method development. All authors reviewed and edited the manuscript. H.L. and J.N. designed and conceived the initial project. J. N. and J. S. obtained the funding for this study.

Conflict of Interests

The authors declare that they have no conflicts of interest.

References

1. Nielsen, J. & Keasling, J. D. Engineering Cellular Metabolism. *Cell* **164**, 1185–1197 (2016).
2. Jullesson, D., David, F., Pfleger, B. & Nielsen, J. Impact of synthetic biology and metabolic engineering on industrial production of fine chemicals. *Biotechnology Advances* **33**, (2015).
3. Ro, D. K. *et al.* Production of the antimalarial drug precursor artemisinic acid in engineered yeast. *Nature* **440**, (2006).
4. Galanie, S., Thodey, K., Trenchard, I. J., Interrante, M. F. & Smolke, C. D. Complete biosynthesis of opioids in yeast. *Science (80-.).* **349**, (2015).
5. Choi, Y. J. & Lee, S. Y. Microbial production of short-chain alkanes. *Nature* **502**, (2013).
6. Zhou, Y. J. *et al.* Production of fatty acid-derived oleochemicals and biofuels by synthetic yeast cell factories. *Nat. Commun.* **7**, (2016).
7. Jin, H. P., Kwang, H. L., Tae, Y. K. & Sang, Y. L. Metabolic engineering of Escherichia coli for the production of L-valine based on transcriptome analysis and in silico gene knockout simulation. *Proc. Natl. Acad. Sci. U. S. A.* **104**, (2007).
8. Wei, Y., Bergenholm, D., Gossing, M., Siewers, V. & Nielsen, J. Expression of cocoa genes in Saccharomyces cerevisiae improves cocoa butter production. *Microb. Cell Fact.* **17**, (2018).
9. Otero, J. M. *et al.* Industrial systems biology of Saccharomyces cerevisiae enables novel succinic acid cell factory. *PLoS One* **8**, e54144 (2013).
10. Nielsen, J. Yeast cell factories on the horizon. *Science (80-.).* **349**, (2015).

11. Nielsen, J. Systems Biology of Metabolism. *Annu. Rev. Biochem.* **86**, 245–275 (2017).
12. Choi, H. S., Lee, S. Y., Kim, T. Y. & Woo, H. M. In silico identification of gene amplification targets for improvement of lycopene production. *Appl. Environ. Microbiol.* **76**, 3097–3105 (2010).
13. Alper, H., Jin, Y. S., Moxley, J. F. & Stephanopoulos, G. Identifying gene targets for the metabolic engineering of lycopene biosynthesis in Escherichia coli. *Metab. Eng.* **7**, (2005).
14. Fong, S. S. *et al.* In silico design and adaptive evolution of Escherichia coli for production of lactic acid. *Biotechnol. Bioeng.* **91**, (2005).
15. Suástegui, M. *et al.* Multilevel engineering of the upstream module of aromatic amino acid biosynthesis in Saccharomyces cerevisiae for high production of polymer and drug precursors. *Metab. Eng.* **42**, (2017).
16. Segre, D., Vitkup, D. & Church, G. M. Analysis of optimality in natural and perturbed metabolic networks. *Proc. Natl. Acad. Sci.* **99**, 15112–15117 (2002).
17. Burgard, A. P., Pharkya, P. & Maranas, C. D. OptKnock: A Bilevel Programming Framework for Identifying Gene Knockout Strategies for Microbial Strain Optimization. *Biotechnol. Bioeng.* **84**, (2003).
18. Ranganathan, S., Suthers, P. F. & Maranas, C. D. OptForce: An optimization procedure for identifying all genetic manipulations leading to targeted overproductions. *PLoS Comput. Biol.* (2010). doi:10.1371/journal.pcbi.1000744
19. Chowdhury, A., Zomorrodi, A. R. & Maranas, C. D. k-OptForce: Integrating Kinetics with Flux Balance Analysis for Strain Design. *PLoS Comput. Biol.* (2014). doi:10.1371/journal.pcbi.1003487
20. Khodayari, A. & Maranas, C. D. A genome-scale Escherichia coli kinetic metabolic model k-ecoli457 satisfying flux data for multiple mutant strains. *Nat. Commun.* **7**, (2016).
21. Domenzain, I. *et al.* Reconstruction of a catalogue of genome-scale metabolic models with enzymatic constraints using GECKO 2.0. *bioRxiv* 2021.03.05.433259 (2021). doi:10.1101/2021.03.05.433259
22. Orth, J. D., Thiele, I. & Palsson, B. Ø. What is flux balance analysis? *Nat. Biotechnol.* **28**, 245–248 (2010).
23. Savinell, J. M. & Palsson, B. O. Network analysis of intermediary metabolism using linear optimization. I. Development of mathematical formalism. *J. Theor. Biol.* **154**, (1992).
24. Vazquez, A. *et al.* Impact of the solvent capacity constraint on E. coli metabolism. *BMC Syst. Biol.* **2**, 7 (2008).
25. Vazquez, A., Liu, J., Zhou, Y. & Oltvai, Z. N. Catabolic efficiency of aerobic glycolysis: The Warburg effect revisited. *BMC Syst. Biol.* **4**, (2010).

26. Nilsson, A. & Nielsen, J. Metabolic Trade-offs in Yeast are Caused by F1F0-ATP synthase. *Sci. Rep.* **6**, 1–11 (2016).
27. Adadi, R., Volkmer, B., Milo, R., Heinemann, M. & Shlomi, T. Prediction of microbial growth rate versus biomass yield by a metabolic network with kinetic parameters. *PLoS Comput. Biol.* **8**, (2012).
28. Chen, Y. & Nielsen, J. Yeast has evolved to minimize protein resource cost for synthesizing amino acids. *Proc. Natl. Acad. Sci.* **119**, e2114622119 (2022).
29. Ishchuk, O. P. *et al.* Genome-scale modeling drives 70-fold improvement of intracellular heme production in Saccharomyces cerevisiae. *Proc. Natl. Acad. Sci.* **119**, e2108245119 (2022).
30. Hassing, E. J., de Groot, P. A., Marquenie, V. R., Pronk, J. T. & Daran, J. M. G. Connecting central carbon and aromatic amino acid metabolisms to improve de novo 2-phenylethanol production in Saccharomyces cerevisiae. *Metab. Eng.* **56**, (2019).
31. Larroude, M., Nicaud, J. M. & Rossignol, T. Yarrowia lipolytica chassis strains engineered to produce aromatic amino acids via the shikimate pathway. *Microb. Biotechnol.* (2020). doi:10.1111/1751-7915.13745
32. Rajkumar, A. S. & Morrissey, J. P. Rational engineering of Kluyveromyces marxianus to create a chassis for the production of aromatic products. *Microb. Cell Fact.* **19**, (2020).
33. Qin, J. *et al.* Engineering yeast metabolism for the discovery and production of polyamines and polyamine analogues. *Nat. Catal.* (2021). doi:10.1038/s41929-021-00631-z
34. Westfall, P. J. *et al.* Production of amorphadiene in yeast, and its conversion to dihydroartemisinic acid, precursor to the antimalarial agent artemisinin. *Proc. Natl. Acad. Sci. U. S. A.* **109**, (2012).
35. Zhang, G. *et al.* Refactoring β-Amyrin synthesis in Saccharomyces cerevisiae. *AIChE J.* **61**, (2015).
36. López, J. *et al.* Production of β-ionone by combined expression of carotenogenic and plant CCD1 genes in Saccharomyces cerevisiae. *Microb. Cell Fact.* **14**, (2015).
37. Srinivasan, P. & Smolke, C. D. Engineering a microbial biosynthesis platform for de novo production of tropane alkaloids. *Nat. Commun.* **10**, (2019).
38. Curran, K. A., Leavitt, J. M., Karim, A. S. & Alper, H. S. Metabolic engineering of muconic acid production in Saccharomyces cerevisiae. *Metab. Eng.* **15**, (2013).
39. Tippmann, S., Scalcinati, G., Siewers, V. & Nielsen, J. Production of farnesene and santalene by Saccharomyces cerevisiae using fed-batch cultivations with RQ-controlled feed. *Biotechnol. Bioeng.* **113**, 72–81 (2016).
40. Jiang, G. Z. *et al.* Manipulation of GES and ERG20 for geraniol overproduction in Saccharomyces cerevisiae. *Metab. Eng.* **41**, (2017).

41. Tang, L. *et al.* Three-pathway combination for glutathione biosynthesis in Saccharomyces cerevisiae. *Microb. Cell Fact.* **14**, (2015).
42. Reifenrath, M. & Boles, E. Engineering of hydroxymandelate synthases and the aromatic amino acid pathway enables de novo biosynthesis of mandelic and 4-hydroxymandelic acid with Saccharomyces cerevisiae. *Metab. Eng.* **45**, (2018).
43. Zelle, R. M. *et al.* Malic acid production by Saccharomyces cerevisiae: Engineering of pyruvate carboxylation, oxaloacetate reduction, and malate export. *Appl. Environ. Microbiol.* **74**, (2008).
44. Zhou, Y. J. *et al.* Modular pathway engineering of diterpenoid synthases and the mevalonic acid pathway for miltiradiene production. *J. Am. Chem. Soc.* **134**, (2012).
45. Meng, X. *et al.* Metabolic engineering Saccharomyces cerevisiae for de novo production of the sesquiterpenoid (+)-nootkatone. *Microb. Cell Fact.* **19**, (2020).
46. Qin, J. *et al.* Modular pathway rewiring of Saccharomyces cerevisiae enables high-level production of L-ornithine. *Nat. Commun.* **6**, (2015).
47. Wang, Z., Gao, C., Wang, Q., Liang, Q. & Qi, Q. Production of pyruvate in Saccharomyces cerevisiae through adaptive evolution and rational cofactor metabolic engineering. *Biochem. Eng. J.* **67**, (2012).
48. Ng, C. Y., Jung, M. Y., Lee, J. & Oh, M. K. Production of 2,3-butanediol in Saccharomyces cerevisiae by in silico aided metabolic engineering. *Microb. Cell Fact.* **11**, (2012).
49. Li, T. *et al.* Metabolic Engineering of Saccharomyces cerevisiae to Overproduce Squalene. *J. Agric. Food Chem.* **68**, (2020).
50. Chen, H. *et al.* High production of valencene in Saccharomyces cerevisiae through metabolic engineering. *Microb. Cell Fact.* **18**, (2019).
51. Pyne, M. E. *et al.* A yeast platform for high-level synthesis of tetrahydroisoquinoline alkaloids. *Nat. Commun.* **11**, (2020).
52. McKeague, M., Wang, Y. H., Cravens, A., Win, M. N. & Smolke, C. D. Engineering a microbial platform for de novo biosynthesis of diverse methylxanthines. *Metab. Eng.* **38**, (2016).
53. Jakočiūnas, T. *et al.* Programmable polyketide biosynthesis platform for production of aromatic compounds in yeast. *Synth. Syst. Biotechnol.* **5**, (2020).
54. Farhi, M. *et al.* Harnessing yeast subcellular compartments for the production of plant terpenoids. *Metab. Eng.* **13**, (2011).
55. der Maaten, L. & Hinton, G. Visualizing data using t-SNE. *J. Mach. Learn. Res.* **9**, (2008).
56. Yu, T. *et al.* Reprogramming Yeast Metabolism from Alcoholic Fermentation to Lipogenesis. *Cell* **174**, (2018).

Main Figures

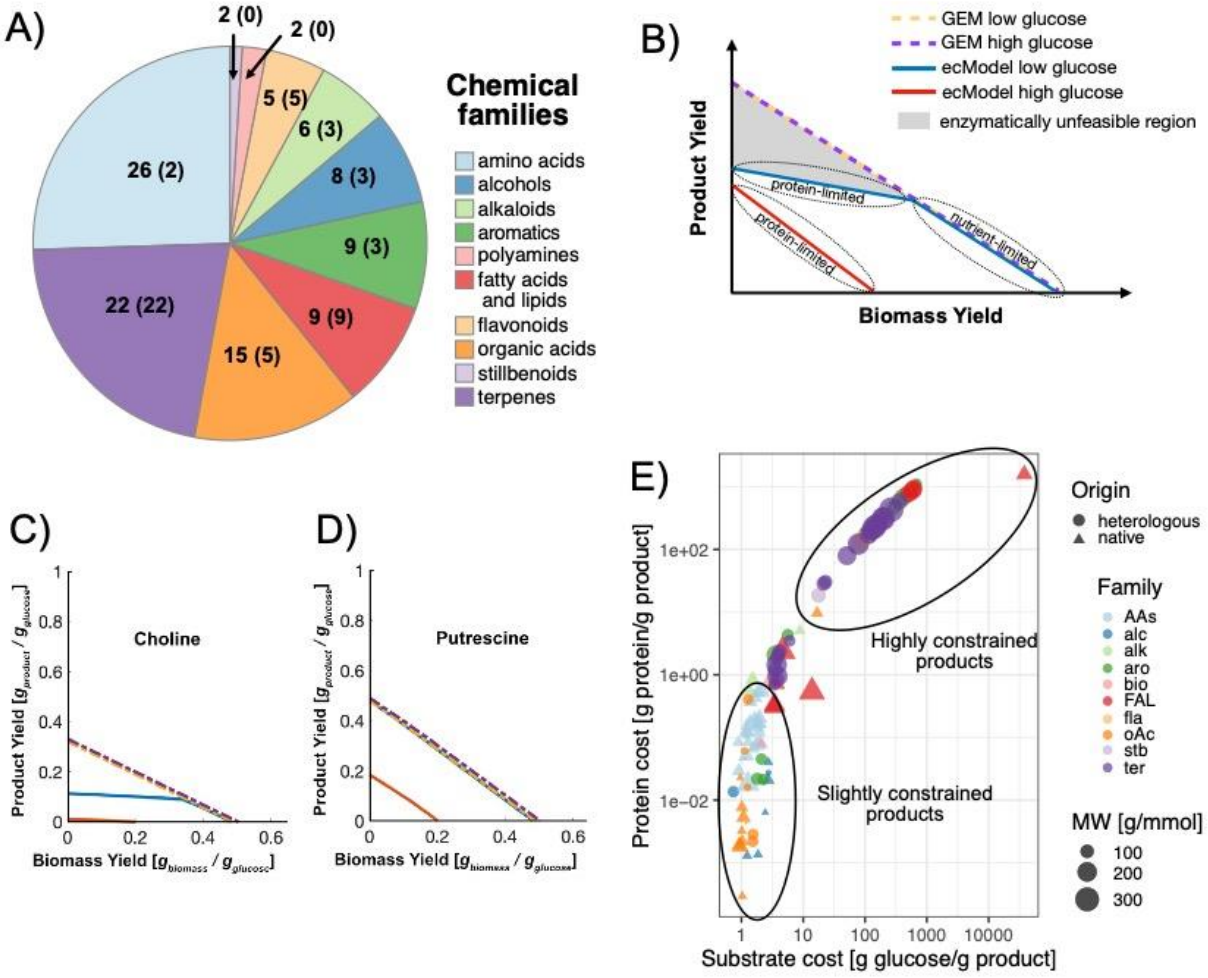


Figure 1.- Exploration of chemical production in yeast using enzyme-constrained metabolic modeling. A) Chemical classification of 102 chemicals for in silico prediction. Numbers within parenthesis indicate number of native products in the different families, those outside the parenthesis indicate the total number of products in the family. B) Production landscape predicted by a metabolic model with and without enzyme constraints at low and high glucose uptake levels. C) Production yield plot for the highly protein-constrained product choline. D) Production yield plot for the substrate-limited putrescine. E) Predicted substrate and protein cost of chemical production in yeast. Product origin, chemical classification and molecular weights are indicated by the characteristics of the 102 markers.

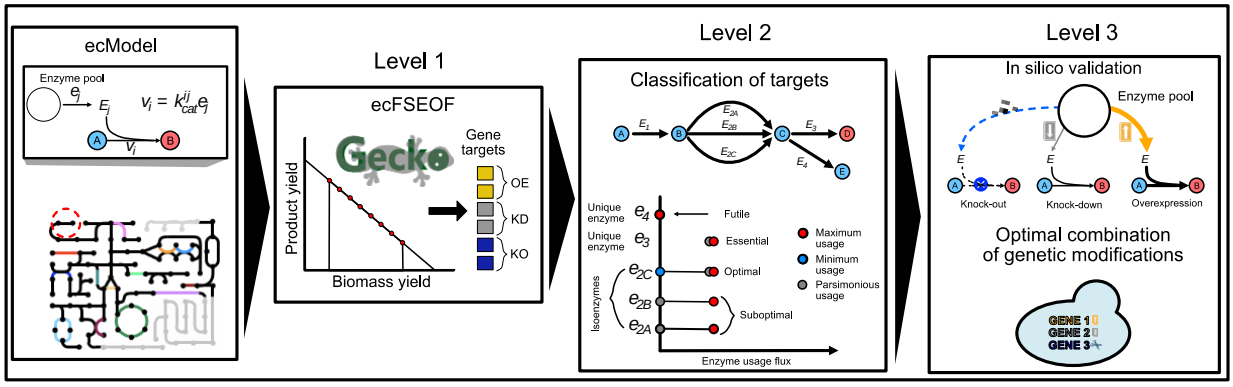


Figure 2.- Prediction of metabolic engineering targets with ecFactory. A metabolic model with enzyme constraints is used for (1) prediction of gene targets for rewiring flux towards increased production. (2) Gene targets are classified and filtered according to enzymatic efficiency and connectivity. (3) A minimal combination of targets for sustaining optimal production levels is obtained.

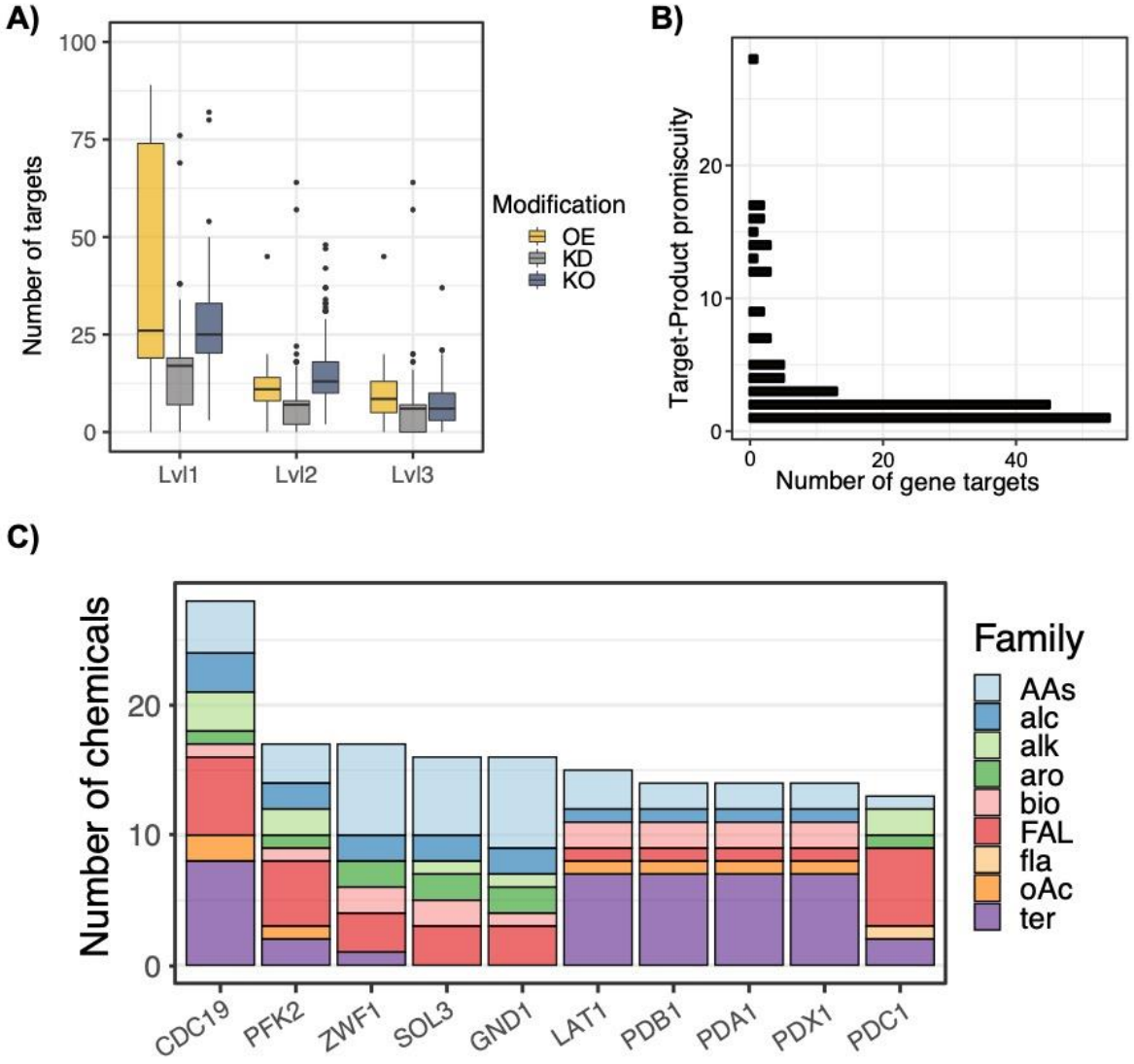


Figure 3.- Prediction of gene engineering targets for increased production of 102 chemicals in yeast. A) Distribution of the number of gene targets per product predicted at different steps in the ecFactory pipeline. Level1, FSEOF-based prediction; Level2, filtering by enzyme characteristics; Level3, obtention of minimal set of targets for optimal production. B) Distribution of product specificity of gene targets across 102 chemicals. C) Representation of the presence of the top 10 most common predicted overexpression targets across products and families.

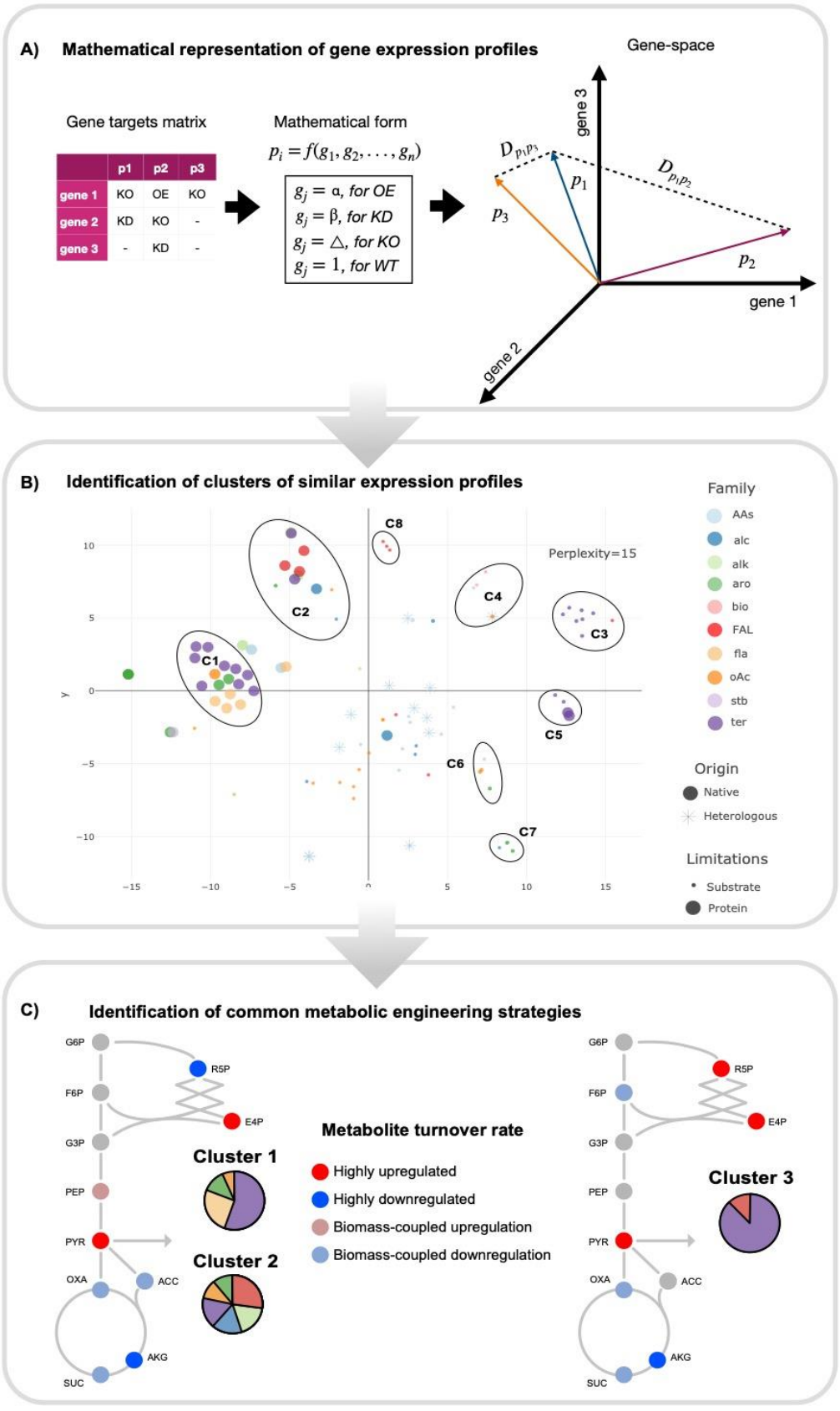


Figure 4.- Model-driven design of platform strains for diverse chemical production. A) representation of gene targets for optimal production as mathematical vectors. B) Identification of clusters of products with similarities in their predicted engineering targets using t-SNE. Chemical families are indicated as AA, amino acids; Alc, alcohols; Alk, alkaloids; Aro, aromatics; Bio, bioamines; FAL, fatty acids and lipids; fla, flavonoids; oAc, organic acids; stb, stillbenoids; ter, terpenes. C) FBA predicts cluster-specific metabolic rewiring strategies. Fold-change in turnover rate of the main metabolic precursors, compared to wild-type, necessary for optimal production of the products in clusters 1, 2 and 3.

carbon metabolism

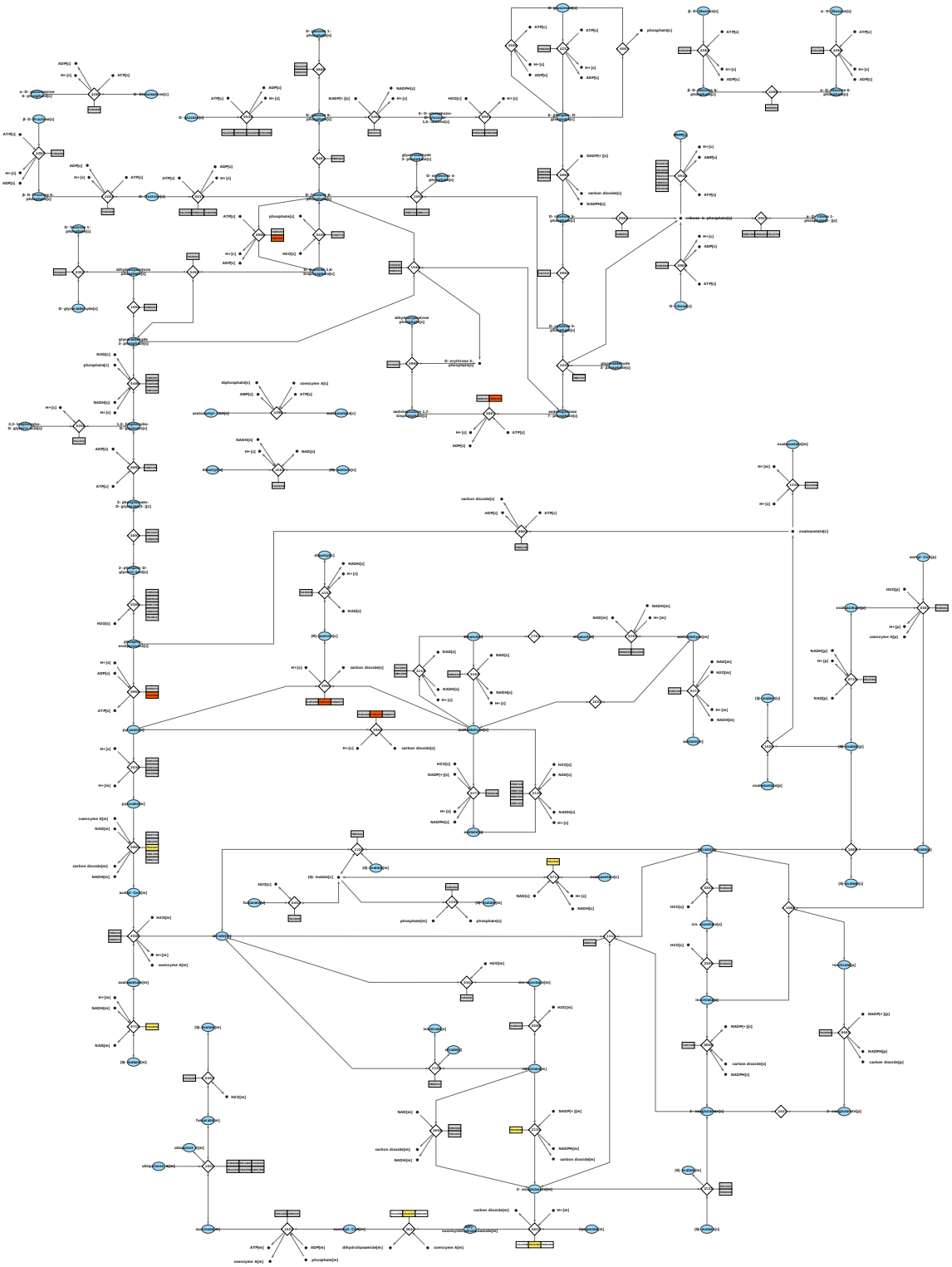


Figure 5.- Map of *S. cerevisiae's* central carbon metabolism from metabolic atlas. Gene targets for increased production of the terpene patchoulol are shown in red, for overexpressions; yellow for down-regulated targets; and gene for predicted gene deletions.

Paper VI:

A novel yeast hybrid modeling framework integrating Boolean and enzyme-constrained networks enables exploration of the interplay between signaling and metabolism

Österberg, L., Domenzain, I., Münch, J., Nielsen, J., Hohmann, S., & Cvijovic, M.

PLOS Computational Biology, 2021

PLOS COMPUTATIONAL BIOLOGY

RESEARCH ARTICLE

A novel yeast hybrid modeling framework integrating Boolean and enzyme-constrained networks enables exploration of the interplay between signaling and metabolism

Linnea Österberg[1,2,3], Iván Domenzain[3,4], Julia Münch[1,2], Jens Nielsen[3,4,5], Stefan Hohmann[3], Marija Cvijovic[1,2]*

1 Department of Mathematical Sciences, University of Gothenburg, Gothenburg, Sweden, **2** Department of Mathematical Sciences, Chalmers University of Technology, Gothenburg, Sweden, **3** Department of Biology and Biological Engineering, Chalmers University of Technology, Gothenburg, Sweden, **4** Novo Nordisk Foundation Center for Biosustainability, Chalmers University of Technology, Gothenburg, Sweden, **5** BioInnovation Institute, Copenhagen, Denmark

* marija.cvijovic@chalmers.se

Citation: Österberg L, Domenzain I, Münch J, Nielsen J, Hohmann S, Cvijovic M (2021) A novel yeast hybrid modeling framework integrating Boolean and enzyme-constrained networks enables exploration of the interplay between signaling and metabolism. PLoS Comput Biol 17(4): e1008891. https://doi.org/10.1371/journal.pcbi.1008891

Editor: Attila Csikász-Nagy, King's College London, UNITED KINGDOM

Received: September 9, 2020

Accepted: March 18, 2021

Published: April 9, 2021

Peer Review History: PLOS recognizes the benefits of transparency in the peer review process; therefore, we enable the publication of all of the content of peer review and author responses alongside final, published articles. The editorial history of this article is available here: https://doi.org/10.1371/journal.pcbi.1008891

Data Availability Statement: All relevant data are within the manuscript and its Supporting Information files. The model, code, and datasets

Abstract

The interplay between nutrient-induced signaling and metabolism plays an important role in maintaining homeostasis and its malfunction has been implicated in many different human diseases such as obesity, type 2 diabetes, cancer, and neurological disorders. Therefore, unraveling the role of nutrients as signaling molecules and metabolites together with their interconnectivity may provide a deeper understanding of how these conditions occur. Both signaling and metabolism have been extensively studied using various systems biology approaches. However, they are mainly studied individually and in addition, current models lack both the complexity of the dynamics and the effects of the crosstalk in the signaling system. To gain a better understanding of the interconnectivity between nutrient signaling and metabolism in yeast cells, we developed a hybrid model, combining a Boolean module, describing the main pathways of glucose and nitrogen signaling, and an enzyme-constrained model accounting for the central carbon metabolism of *Saccharomyces cerevisiae*, using a regulatory network as a link. The resulting hybrid model was able to capture a diverse utalization of isoenzymes and to our knowledge outperforms constraint-based models in the prediction of individual enzymes for both respiratory and mixed metabolism. The model showed that during fermentation, enzyme utilization has a major contribution in governing protein allocation, while in low glucose conditions robustness and control are prioritized. In addition, the model was capable of reproducing the regulatory effects that are associated with the Crabtree effect and glucose repression, as well as regulatory effects associated with lifespan increase during caloric restriction. Overall, we show that our hybrid model provides a comprehensive framework for the study of the non-trivial effects of the interplay between signaling and metabolism, suggesting connections between the Snf1 signaling pathways and processes that have been related to chronological lifespan of yeast cells.

used for this study can be found in the GitHub repository YeastHybridModelingFramework https://github.com/cvijoviclab/YeastHybridModelingFramework.

Funding: This work was supported by the Swedish Research Council (VR2016-03744) to SH, Swedish Foundation for Strategic Research (Grant Nr. FFL15-0238) to MC and European Union's Horizon 2020 research and innovation program, project CHASSY (grant agreement 720824) to ID. Part of this work was funded by the Novo Nordisk Foundation (grant no. NNF10CC1016517) and the Knut and Alice Wallenberg Foundation to JN. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing interests: The authors have declared that no competing interests exist.

Author summary

Elucidating the complex relationship between nutrient-induced signaling and metabolism represents a key in understanding the onset of many different human diseases like obesity, type 3 diabetes, cancer, and many neurological disorders. In this work we proposed a hybrid modeling approach, combining Boolean representation of signaling pathways, like Snf1, TORC1, and PKA with the enzyme constrained model of metabolism linking them via the regulatory network. This allowed us to improve individual model predictions and elucidate how single components in the dynamic signaling layer affect steady-state metabolism. The model has been tested under respiration and fermentation, revealing novel connections and further reproducing the regulatory effects that are associated with the Crabtree effect and glucose repression. Finally, we show a connection between Snf1 signaling and chronological lifespan.

Introduction

Biological systems are of complex nature comprising numerous dynamical processes and networks on different functional, spatial and temporal levels, while being highly interconnected [1]. The field of systems biology faces the great challenge of elucidating how these interconnected systems work both separately and together to prime organisms for survival. One such phenomenon is the cells' ability to sense and respond to environmental conditions such as nutrient availability. To coordinate cellular metabolism and strategize, the cell needs an exact perception of the dynamics of intra- and extracellular metabolites [2]. Simultaneously, nutrient-induced signaling plays a pivotal role in numerous human diseases like obesity, type 2 diabetes, cancer and aging [3–6]. Therefore, unraveling the role of nutrients as signaling molecules and metabolites as well as their interconnectivity may provide a deeper understanding of how these conditions occur.

Yeast has long been used as a model organism for studying nutrient-induced signaling [7]. Two major classes of nutrients include carbon and nitrogen. Carbon-induced signaling acts mainly through the PKA and SNF1 pathway while nitrogen-induced signaling acts through the mTOR pathway. The PKA pathway plays a major role in regulating growth by inducing ribosome biogenesis genes and inhibiting stress response genes [8]. The SNF1 pathway is mainly active in low glucose conditions where it promotes respiratory metabolism, glycogen accumulation, gluconeogenesis, and utilization of alternative carbon sources but it also controls cellular developmental processes such as meiosis and aging [7, 9, 10]. The strongly conserved TORC1 pathway plays a crucial role in promoting anabolic processes and cell growth in response to nitrogen availability [8]. Active TORC1 induces ribosomal protein and ribosome biogenesis gene expression [11, 12] and represses transcription of genes containing STR and PDS elements in their promoter region [11]. Even though Snf1, TORC1, and PKA pathways belong to the most well-studied pathways [2], there is still a lack of understanding both in the dynamics and the interactions leading to change in gene expression. It has been shown, that glucose signaling is related to metabolism however the nature of this relationship remains unknown [13]. Numerous crosstalk mechanisms between these pathways have been described [14], and depending on their activity, they may influence the overall effect of the signaling process and thus the interaction with the metabolism [15]. To better understand the impact of cell signaling on metabolism, a systems biology approach is often implemented [16].

Typically, Boolean models have been developed to study the crosstalk between the Snf1 pathway and the Snf3/Rgt2 pathway [17] as well as the Snf1, cAMP-PKA, and Rgt2/Snf3 pathways [15]. In mammalian cells, Boolean models have been used to evaluate the conflicting hypothesis of the regulation of the mTOR pathway [18] and to study crosstalk between mTOR and MAPK signaling pathways [19]. Since, signaling systems are not always strictly Boolean in its nature, where location, combinations of post-translational modifications as well as other interaction play a role, alternative Boolean frameworks for handling these complex interactions have been developed [15, 20]. In contrast, metabolism, also in itself a complex process, is often studied using Flux Balance Analysis (FBA), which enables prediction of biochemical reaction fluxes, cellular growth on different environments, and gene essentiality even for genome-scale metabolic models [21–23]. A major limitation of the use of GEMs together with FBA is the high variability of flux distributions for a given cellular objective [24], as FBA solves largely underdetermined linear systems through optimization methods. To overcome this problem, experimentally measured exchange fluxes (uptake of nutrients and secretion of byproducts) are incorporated as numerical constraints, however, such measurements are not readily available for a wide variety of organisms and growth conditions.

The concept of enzyme capacity constraints has been incorporated into FBA to reduce the phenotypic solution space (i.e. exclusion of flux distributions that are not biologically meaningful) and diminish its dependency on condition-dependent exchange fluxes datasets [25–30]. Notably, a method to account for enzyme constraints, genome-scale models using kinetics and omics (GECKO; Sánchez et al., 2017) has been developed. GECKO incorporates constraints on metabolic fluxes given by the maximum activity of enzymes, which are also constrained by a limited pool of protein in the cell. This method has refined predictions for growth on diverse environments, cellular response to genetic perturbations, and even predicted the Crabtree effect in *S. cerevisiae*'s metabolism, but also proven to be a helpful tool for probing protein allocation and enabled the integration of condition-dependent absolute proteomics data into metabolic networks [28, 30].

Following the holistic view of systems biology, hybrid models allow us to take the next step and combine different formalisms to study the interconnectivity and crosstalk spanning different scales and/or systems. For example, to quantify the contribution of the regulatory constraints of an *Escherichia coli* genome-scale model, a steady-state regulatory flux balance analysis (SR-FBA) has been developed [31]. Furthermore, the diauxic shift in *S. cerevisiae* has been studied by the CoRegFlux workflow, integrating metabolic models and gene regulatory networks [32]. To bypass the need for kinetic parameters, a FlexFlux tool has been developed where metabolic flux analyses using FBA have been constrained with steady-state values resulting from the regulatory network [33]. This strategy has also been used in a hybrid model of *Mycobacterium tuberculosis* where the gene regulatory network was used to constrain the metabolic model to study the adaptation to the intra-host hypoxic environment [34]. However, to further study the impact of signaling on metabolism, the complexity of the signaling systems itself and the crosstalk between interacting pathways need to be represented coherently.

To better understand the complex relationship between metabolism and signaling pathways, we created a hybrid model consisting of a Boolean module integrating the PKA, TORC1, and the Snf1 pathways as well as the known crosstalk between, together with an enzyme-constrained module of *S. cerevisiae's* central carbon and energy metabolism (Fig 1). The backbone of the presented model is a framework for utilizing the complex Boolean representation of large-scale signaling systems to further constrain an enzyme-constrained model (ecModel) of the central carbon metabolism. With the glucose level as input, transcription factor activities, resulting from the Boolean module are mapped to a regulatory network. The bounds of the solution space, calculated using enzyme usage variability analysis on the genes affected by the

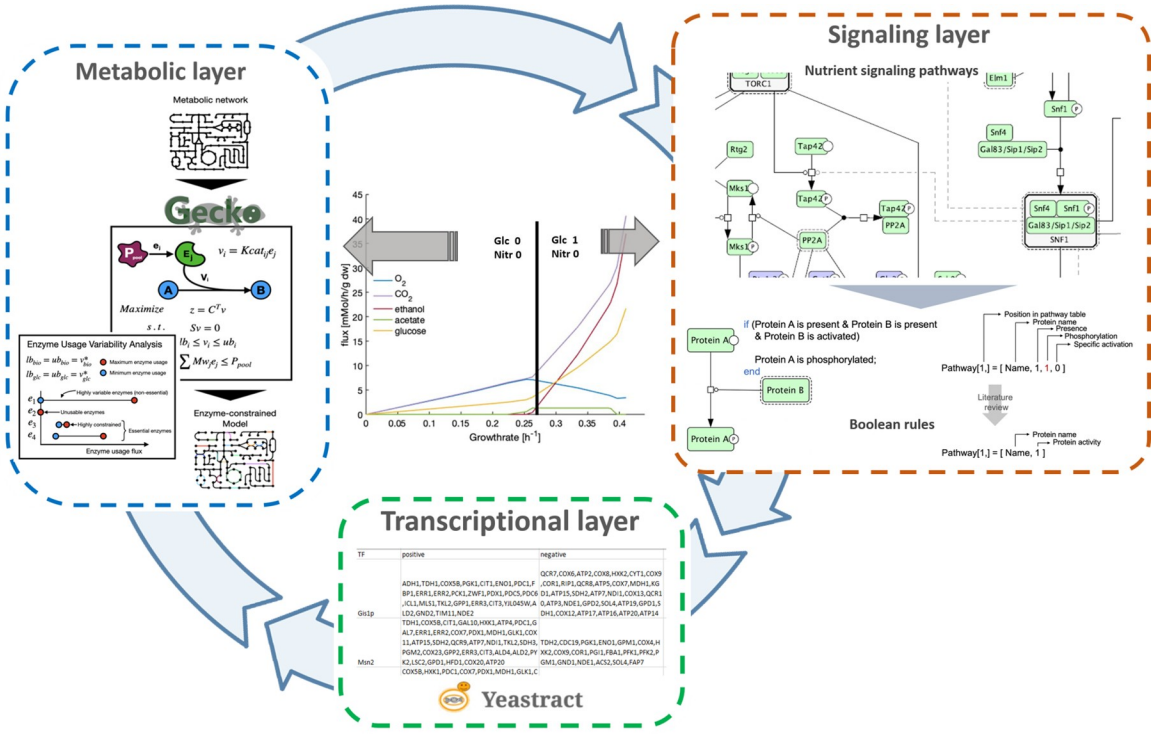


Fig 1. Schematic representation of the hybrid model. The hybrid model consists of a vector-based Boolean module of nutrient signaling and an enzyme constrained module of the central carbon metabolism. The Boolean module is a dynamic module including Snf1, PKA, and TORC1 pathway as well as crosstalk between them. The dynamic module reaches a steady-state and the activity of the transcription factors acts as input in a regulatory network constraining the enzyme constraint model of the central carbon metabolism. The solution is used to determine the activity of the Boolean input.

https://doi.org/10.1371/journal.pcbi.1008891.g001

transcription factors, are altered depending on up- or down- regulation and used to constrain an ecModel of the central carbon metabolism (for details see Method section). The predictions of protein allocation, at the individual enzyme level for respiratory and fermentative conditions, are considerably improved by the incorporation of the regulatory layer into the hybrid model, in comparison with its pure enzyme-constrained counterpart. Moreover, the predicted enzyme usage profiles display a diversified utilization of isoenzymes, which is supported by proteomics data, but previous constraint-based methods have failed to capture. Additionally, the proposed hybrid model is capable of reproducing the regulatory effects that are associated with the Crabtree effect and glucose repression and have further showed a connection between glucose signaling and chronological lifespan by the regulation of NDE and NDI usage in respiring conditions. Finally, the model showed that during fermentation, enzyme utilization is the more important factor governing protein allocation, while in low glucose conditions robustness and control are prioritized.

Results

Implemented Boolean signaling network reproduces the general dynamics caused by glucose and nitrogen addition to starved cells

To verify the constructed Boolean model of nutrient-induced signaling pathways (Fig 2), cells were simulated from nitrogen- and glucose-starved conditions to nutrient-rich conditions. We also simulated the model from nutrient-rich conditions to nutrient depletion. The

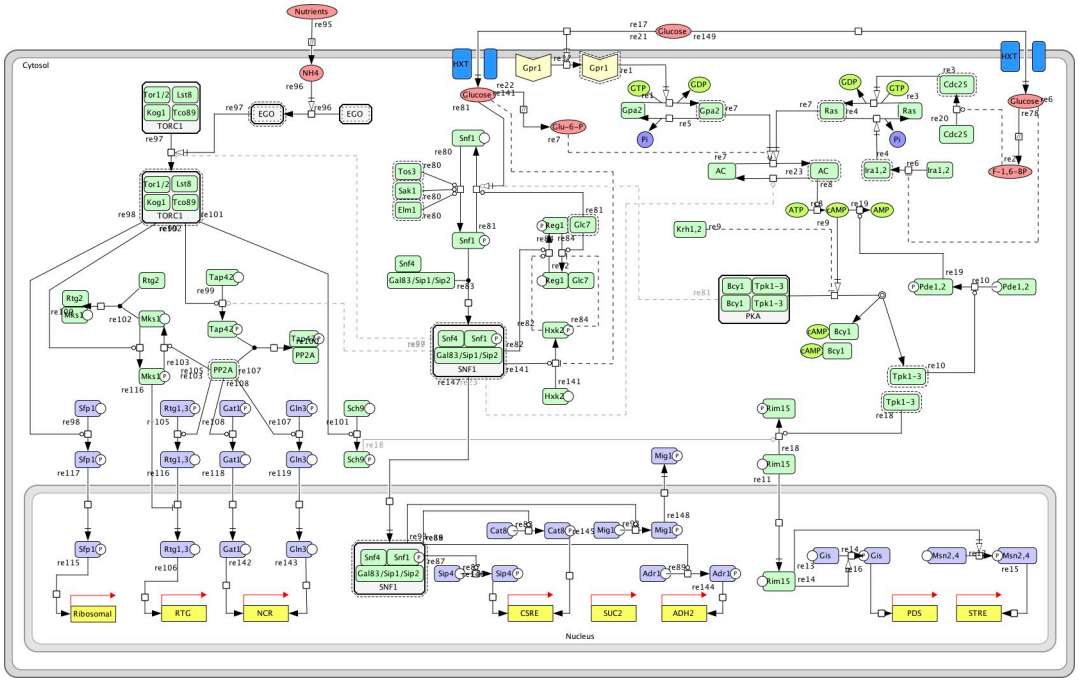


Fig 2. The Boolean module is a dynamic module including Snf1, PKA, and TOR pathway as well as crosstalk between them. Crosstalk events between the pathways are depicted in grey. Unknown mechanisms are represented by dashed lines.

https://doi.org/10.1371/journal.pcbi.1008891.g002

simulated results were compared to the literature concerning both dynamics and steady-state outcome. An in-depth literature review of the known mechanisms and its implementation and interpretation in the model as well as a graphical representation of the simulated dynamics are available in S1 Text, S1 Fig, and S1 Table. The PKA pathway was activated upon glucose abundance via the small G proteins Ras and Gpa2. These proteins, in turn, activated the adenylate cyclase (AC) that induced the processes leading to the activation of the catalytic subunit of PKA. Active PKA phosphorylated and therefore inactivated Rim15, thus the transcription factors Gis1, Msn2, and Msn4 became inactive. Our result pinpointed PKA as the main regulator of Rim15 (for details see S1 Text), while previous experimental studies showed that Sch9 is the major regulator of Rim15 [35]. Further, simulations from high to low nutrient conditions are in agreement with the literature on dynamics and steady-states (S1 Fig). When glucose is depleted Ira becomes active and sequentially Cdc25 gets inactivated which results in Ras inactivation. Simultaneously Gpr1 gets inactivated, turns off Gpa2 relieving the inhibitory effect on Krh activity. This inactivates AC and in turn PKA. Pde and Rim15 get dephosphorylated and Rim15 can phosphorylate Gis1 and Msn2/4.

The SNF1 pathway is active when glucose is limited, while the addition of glucose causes Snf1 inactivation resulting in the activation of the transcriptional repressor Mig1 and the deactivation of Adr1, Cat8, and Sip4. However, the inactivation of Adr1 happened before Snf1 inactivation. This is due to the implemented crosstalk with the PKA pathway, where activated PKA inhibits Adr1 activity [36]. This crosstalk has a similar effect on the dynamics of Adr1 activation in simulations from high to low nutrient conditions. Snf1 is phosphorylated by the upstream kinases when glucose is depleted and then phosphorylates Mig1, Cat8, Sip4, Adr1, and Reg1. Cat8 and Sip1 become active while Adr1 activation occurs two iterations later due to the crosstalk implementation between the PKA and SNF1 pathway (for details see S1 Text).

First, when PKA gets inactive the inhibitory effect it has on Adr1 releases. Mig1 gets inactivated and the phosphorylation of Reg1 activates Glc7.

Nutrient availability activates the TOR complex 1 which in turn phosphorylates Sch9 and Sfp1 resulting in the repression of Rim15 phosphorylation and the expression of ribosomal genes respectively. No change was observed in the activity of PP2A-regulated transcription factors Rtg1, Rtg3, Gat2, and Gln2. However, during the 8^{th} iteration, PP2A was active. In addition, Sch9 was not the main regulator of Rim15 activity in our simulations since PKA was activated before Sch9 and acted independently to regulate Rim15, either due to a gap in the model or a lack of complexity in our understanding of the signaling system (S1 Text). When glucose is depleted the EGO complex loses activity which transmits to the TORC1 complex and in turn to Sch9 and Sfp1.

The Boolean model reveals interconnectivity and knowledge gaps in nutrient signaling pathways

To further investigate the impact of nutrient conditions on the crosstalk between pathways in the Boolean model, knockouts of main components of each pathway (Snf1, Reg1, Tpk1-3, and Tor1,2) were simulated and compared to the wildtype in glc|nitr = 1|1 and glc|nitr = 0|0 (S2 Fig). In nutrient-depleted conditions, only the Snf1 knockout had a significant impact. In the Snf1 pathway, Snf1 knockout affected all downstream targets leading to a transcription factor activity pattern that is usually observed in wildtype strains when glucose is available [7]. It has been previously described that the phenotype of Snf1 mutants resembles the phenotype observed when the cAMP/PKA pathway is over-activated [37]. Although activation of the adenylate cyclase (AC) could be observed in the simulated knockout, PKA and the downstream targets were inactive due to the activity of the Krh proteins that inhibit PKA if no glucose is present in the Boolean model (S1 Text). The Snf1 mutant showed defects in the TOR pathway upon glucose depletion leading to the activation of the PP2A phosphatase. The resulting activation of NCR and RTG genes and deactivation of ribosomal genes correspond to the phenotype one would expect if glucose but not nitrogen is available [38] thus stressing the role of Snf1 in imparting the glucose state to the other nutrient-signaling pathways.

Under high nutrient availability, the Reg1 knockout showed almost the same effect on the SNF1 and TORC1 pathway as nutrient depletion. Only Adr1 activity was not affected which opposes the observations by Dombek and colleagues [39], that described constitutive ADH2 expression in Reg1 mutant cells (S1 Text).

An almost similar effect on the SNF1 and TORC1 pathways could be observed when Tpk1-3 knockout was simulated. This redundant effect was expected since impaired PKA activity was described to be associated with increased SNF1 activity[40]. Nevertheless, PKA knockout additionally induced Adr1 activation when SNF1-mediated activation could no longer be inhibited by PKA. The PKA knockout simulation showed strong effects on all three simulated pathways and may explain why strains lacking all three Tpk isoenzymes are inviable [41].

The effects of Tor1 and 2 knockouts only affected the TORC1 signaling pathway. The simulated phenotype equaled the phenotype that is expected upon nitrogen depletion and glucose abundance and was therefore similar to the phenotype observed when simulating the Snf1 knockout in nutrient-starved cells. Besides, experimental observations revealed that impairing Tor1 and 2 function results in growth arrest in the early G1 phase of the cell cycle, as well as inhibition of translation initiation which are characteristics of nutrient, depleted cells entering stationary-phase [42]. The fact that inactivation of TORC1 results in the inactivation of Sfp1 that regulates the expression of genes required for ribosomal biogenesis could be an indicator of this observation; however other TORC1-associated signaling mechanisms inducing translation initiation may likely be involved [42].

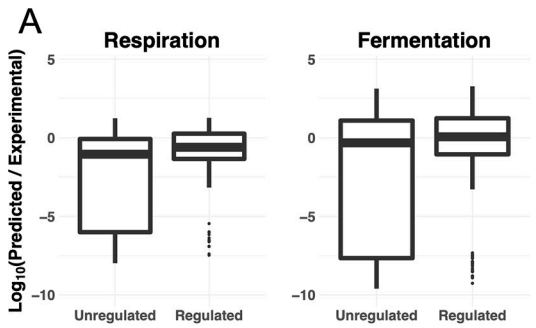
The hybrid model improves protein allocation predictions by showing a diversified use of isoenzymes

To verify the performance of the ecModel layer, predicted exchange fluxes at increasing dilution rates on glucose-limited conditions were compared against experimental data [43] (S3 Fig), predictions showed a median relative error of 9.82% in the whole range of dilution rates from 0 to 0.4 h^{-1}, spanning both respiratory and fermentative metabolic regimes. The hybrid model, including regulation, was further compared with the ecModel in its ability to compute protein demands by comparing the predicted enzyme usages to protein abundance data from the literature, in both respiratory and fermentative conditions [44, 45]. Analysis of results revealed that, in respiration, 40.83% of the proteins in the model are predicted in the same order of magnitude as their experimental values, and 31.66% are predicted with an error between one and two orders of magnitude, yielding an average absolute $\log_{10}$ fold-change between predictions and measurements of 1.55. For the fermentative condition, 65.51% of the proteins are predicted within the same order of magnitude as their experimental measurements, showing an average absolute $\log_{10}$ fold-change of 2.32 (S1 Data and S2 Text). Furthermore, two-sample Kolmogorov-Smirnov tests did not show statistically significant differences between the hybrid model predictions and the available proteomics datasets.

Pathway enrichment analysis of the proteins miss-predicted by more than one order of magnitude by the hybrid model was performed using a hypergeometric distribution test and the Holm-Bonferroni correction method for multiple testing. Results showed that the super-pathway of glucose fermentation was significantly enriched for underpredicted proteins on both respiratory and fermentative conditions (p-value of $1.39\text{x}10^{-7}$ and $7\text{x}10^{-5}$, respectively); additionally, TCA and glyoxylate cycles showed significant enrichment for underpredicted proteins uniquely in fermentation (p-values of $3\text{x}10^{-2}$). On the other hand, the super-pathways of aerobic fermentation and electron transport chain were significantly enriched for overpredicted proteins in the fermentative condition (p-value = $2.85\text{x}10^{-23}$). The pentose phosphate pathway and glucose-6-phosphate biosynthesis showed significant enrichment for underpredicted proteins just in the respiratory condition (p-values of $2.86\text{x}10^{-4}$ and $1.95\text{x}10^{-2}$, respectively). A detailed comparison between the model predictions and in-depth results from the protein predictions are available in S1 Data, S2 Table, and S2 Text.

Comparison with the pure enzyme-constrained model showed that, by adding the regulation layer, prediction of protein demands are improved by more than one order of magnitude, on average, as the aforementioned $\log_{10}$-transformed ratio is reduced from 2.62 to 1.55, in respiration, and from 3.56 to 2.32 for fermentation. This large improvement is predominantly resulting from the utilization of more than one isoform for some reactions in the hybrid model in contrast to a pure ecModel, in which just the most efficient enzyme for a given reaction is used, due to its reliance on optimality principles.

Utilization of isoenzymes was assessed by comparing predicted non-zero enzyme usages, for different isoforms in a given metabolic reaction, to their presence in the datasets for both conditions, returning confusion matrices for the ecModel and hybrid model in each condition (S1 Data). Fig 3 provides a detailed comparison of isoenzymes presence in unregulated and regulated model predictions and proteomics datasets. Predictive performance was then evaluated by computing sensitivity, specificity, precision, accuracy and the Fowlkes-Mallow index, which takes into account all the pair of points in which two clusters of data agree or disagree, approaching the value of one for highly similar clusters [46]. Overall, these metrics revealed that the hybrid model outperforms the ecModel in its ability to predict utilization of expressed isoenzymes in both respiration and fermentation conditions. Further details on predictive performance assessment are shown in Fig 3B.



B

	Respiration		Fermentation	
Metric	ecModel	hybrid	ecModel	hybrid
Sensitivity	0.38	0.66	0.33	0.65
Precision	0.82	0.60	1	1
Accuracy	0.55	0.5	0.35	0.65
FMI	0.57	0.63	0.58	0.80

C

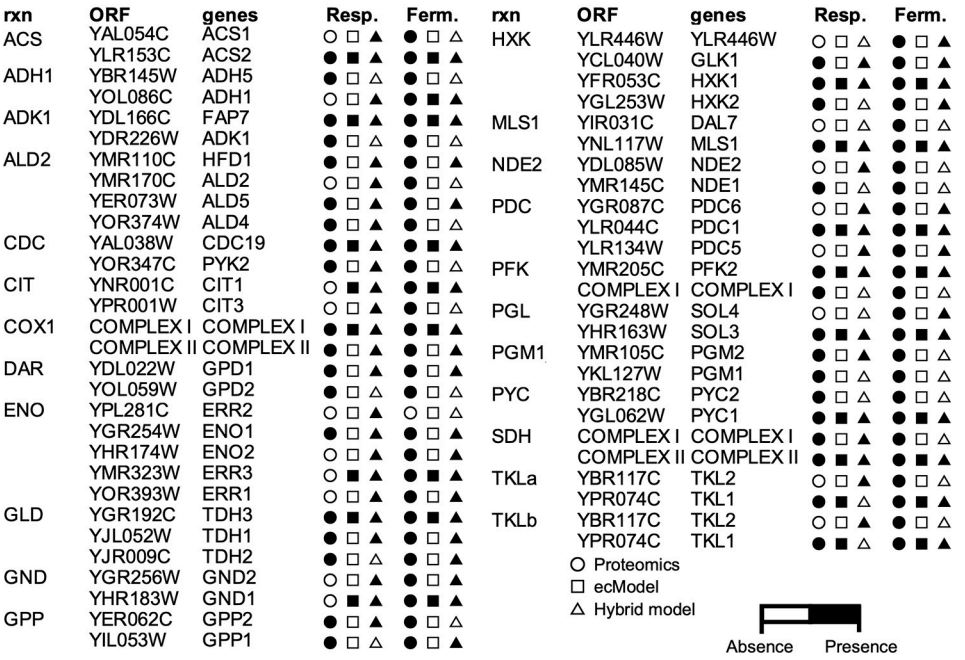


Fig 3. (A) Absolute $\log_{10}$-transformed ratio between predicted and measured protein abundance values in respiration and fermentation for the purely enzyme-constrained and hybrid models. (B) Evaluation of isoenzymes utilization predictions, comparing the ecModel and hybrid model on respiratory and fermentative conditions against experimental data on protein expression (absence/presence). FMI—Fowlkes-Mallows index. (C) Comparison of individual isoenzymes utilization between models' predictions and experimental data. Color indicates presence or absence of a given protein in the predictions of the ecModel, hybrid model and experimental data on protein expression.

https://doi.org/10.1371/journal.pcbi.1008891.g003

The hybrid modeling framework reveals a connection between regulation and chronological aging as well as fundamental strategies of enzyme utilization

To better understand which pathways and reactions are most affected by regulation, the metabolic flux distributions predicted by the hybrid model and the ecModel were compared. Larger flux differences arose for respiratory conditions, in which the average relative change in flux was 1.85 in contrast to 0.46 in fermentation (S2 Data), this result is heavily influenced by the amount of totally activated or deactivated fluxes by the hybrid model, 57 for respiration and 29 for fermentation (Fig 4 and S2 Data). In the ecModels formalism reversible metabolic reactions are split, creating separate reactions for the forward and backward fluxes, thus distributions of

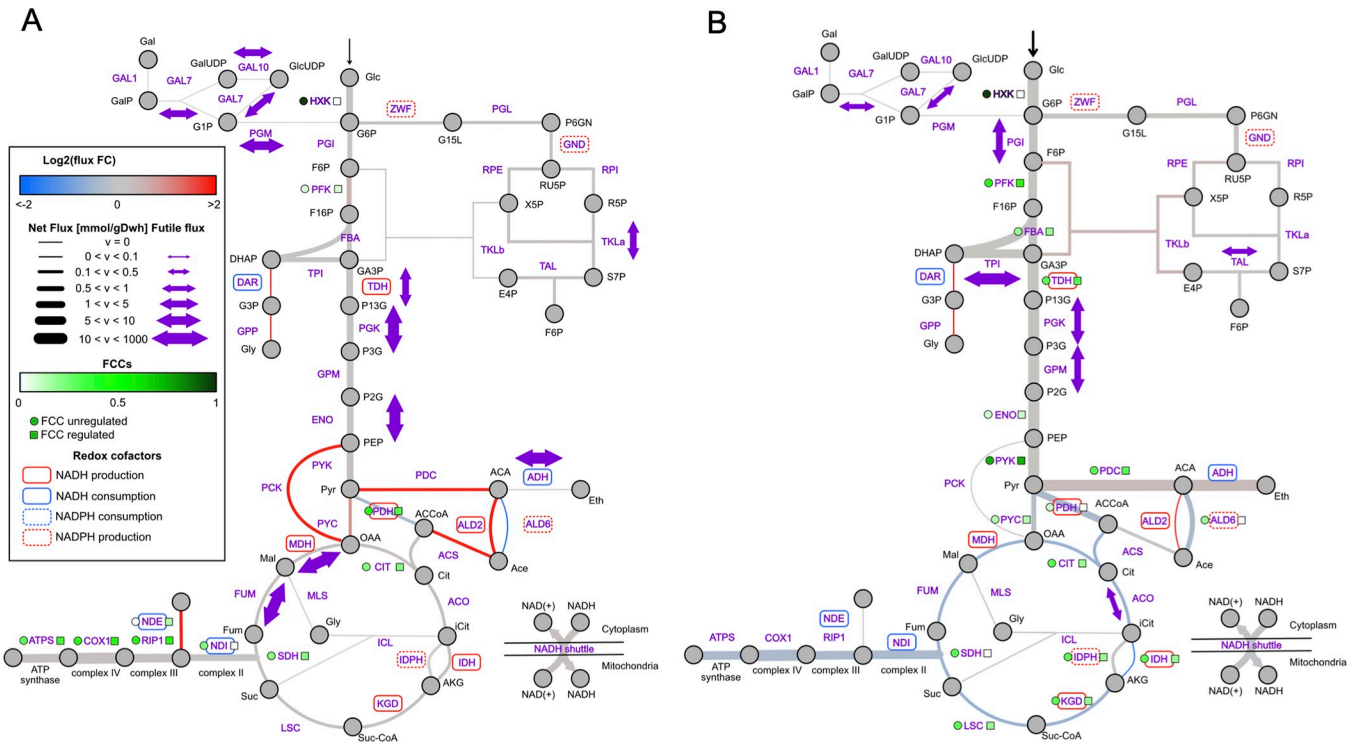


Fig 4. The core reactions in the metabolism under (A) respiration and (B) fermentation are shown. The fluxes are represented by the width of the connectors where dotted lines represent zero flux. The color of the connectors represents the change in flux from the unregulated ecModel compared to the regulated hybrid model. The FCCs are represented in the figure where the unregulated case is depicted by circles and compared to the regulated case depicted in squares.

https://doi.org/10.1371/journal.pcbi.1008891.g004

net metabolic fluxes were also obtained and compared among models and conditions. As some enzymes are upregulated by the hybrid model even to levels that exceed the flux capacity of certain pathways (for a fixed growth rate), futile fluxes are expected to arise across the metabolic network.

Increased exchange fluxes for glucose, oxygen, and acetate were observed in the respiration phase (S2 Data). Additionally, an increase in the overall flux through the pentose phosphate pathway as well as an induced use of NDE instead of only NDI, allowing for the utilization of cytosolic NADH to reduce oxygen demands in the oxidative phosphorylation pathway was detected and has previously been associated with chronological aging [47]. Increased flux on PCK, PDC, ALD2, and ACS, around the pyruvate branching point, led to an overall increased flux through the TCA cycle (Fig 4 and S2 Data). To balance the increased production of AMP by ACS the ADK reaction is also upregulated. Futile fluxes are induced by regulation in galactose metabolism (GAL7 and GAL10), lower glycolysis (TDH, PGK, and ENO), TCA cycle (FUM and MDH), as well as TKLa and PGM in the pentose phosphate pathway and ADH (Fig 4 and S2 Data) in respiration.

In the fermentation state, futile fluxes also occur in galactose metabolism (GAL7) as well as glycolysis (PGI, PGK, TPI, and GPM), TAL1 in the pentose phosphate pathway, and ACO in the TCA cycle (S4 Fig and S2 Data). Down-regulation of oxidative metabolism increased uptake of glucose, and increased flux through glycolysis was observed, which is consistent with the changes that have been attributed to glucose-induced repression during the long term Crabtree effect [48] (Fig 4).

The control exerted by each enzyme on the global glucose uptake rate was investigated through the calculation of flux control coefficients (FCCs), allowing comparison of the

distribution of metabolic control between the pure enzyme-constrained and hybrid model. In both conditions the FCCs obtained for hexokinases by the ecModel (YLR446W for respiration and HXK1 in fermentation) showed a value equal to 1, the highest value in their respective distributions, indicating that the overall glucose uptake rate is mostly governed by the activity of this enzymatic reaction step. In contrast, the constraints applied by the hybrid model distribute the control over the glucose uptake flux in a more even way across different enzymes and pathways, yielding FCCs of 0 for the different HXK isoforms in both metabolic regimes.

As a general trend, more FCCs with a high value (FCC>0.05) are obtained for fermentative conditions than for respiration, despite the use of the ecModel or hybrid model (Fig 4 and S3 Data). In the respiratory condition, the highest FCCs are concentrated in the oxidative phosphorylation pathway as well as around the branching point of pyruvate and PFK in glycolysis, whose activity is related to the connections between glycolysis and PP pathway. Moreover, the absence of glucose uptake control by lower glycolytic enzymes, and the prevalence of a non-zero FCC for PFK in both the ecModel and the hybrid model agrees with experimental evidence for mouse cell-lines in respiratory conditions [49]. For the fermentative condition, the highest FCCs are concentrated in the TCA cycle. Similarly to the respiration case, non-zero FCCs are present in the reaction steps surrounding the connecting points of different pathways, such as PFK, FBA, and TDH connecting glycolysis with the pentose phosphate pathway and reactions around pyruvate, which connect glycolysis with fermentation and the TCA cycle (Fig 4), this trend might indicate that in these branching points kinetic control is still a relevant mechanism governing fluxes.

Deletion of the Snf1 in the hybrid model shows the importance of the Snf1 pathway in low glucose conditions and a connection between Snf1 regulation and chronological aging

To investigate how the individual signaling pathways contribute to changes in metabolic fluxes, the main component of each signaling pathway was deleted and flux changes were compared between the wild-type hybrid model and the knockout versions (S4 Data).

The Snf1 deletion was the only deletion showing any effect on the net fluxes in the respiratory condition (S4 Fig) while the Reg1, PKA and TOR deletions showed effects in fermentation conditions, consistent with the deletion experiments done with the Boolean model. The different mutants in fermentation do not induce major changes in net fluxes, however, the enzyme usage profile differs across the different mutants. Notably, the largest changes in terms of futile fluxes were observed in the TPI reaction, repressed in respiration by the Snf1 pathway and activated in fermentative conditions by either the PKA or Reg1 pathways. In respiration, Snf1 is also responsible for the futile fluxes through GPM, PGI and reduces the futile fluxes through FUM, MDH, PGM, and GAL10. The model simulations show a less diverse use of isoenzymes in all knockouts, which is most likely due to the reduction in the complexity of the regulatory layer. Considering the inherent property of flux balance analysis, any reduction in the regulatory network will be closer to the optimal distribution in which just the most efficient isoforms are used.

The Snf1 deletion exhibits an overall decrease in the flux towards respiration and a large decrease in flux through PPP, showing also a relatively strong downregulation of enzymatic steps surrounding pyruvate. The most significant changes are observed in NDE and PCK that are turned off and ALD6 which is turned on, implying that the Snf1 pathway is responsible for changing the acetate production via ALD6 to acetate production via ALD2, resulting in increased production of cytosolic NADH to the expense of the NADPH, which is compensated by increasing the flux through the pentose phosphate pathway as well as the additional use of

NDE. The use of ACS1 is abolished in the Snf1 deletion but not in other deletions, in the same manner as NDE. ACS1 has been shown to be upregulated in long-lived cells exposed to caloric restriction [50]. We compared the expression in our deletion simulations to experimental data of differentially expressed genes having a positive effect on chronological lifespan [51]. Out of 17 differentially expressed genes covered by our model, 6 were significant at a p-value = 0.05. Of those, 5 were also expressed in our model (S4 Data). RPE1 and CIT3 were upregulated in calorically restricted conditions. CIT3 is upregulated in our model by the Snf1 pathway while RPE1 is regulated by two mechanisms, one visible as a differential expression in the regulated WT compared to the ecModel, and one acting through the SNF1 pathway. RPE1 is also upregulated in the long-lived ade4 mutant strain together with the ADH5 gene. In our model ADH5 is not expressed in the ecModel, the hybrid model, or any of the simulated mutants. Cells treated with concentrates of media from cells grown under caloric restriction show an upregulation of PFK2. This is also shown in cells grown with the drug isonicotinamide (INAM) in which also PGK1 and ENO2 are upregulated. In the hybrid model, PGK1 and ENO2 are upregulated compared to the ecModel and none of the mutant strains showed any differential expression from the hybrid model of the WT. PFK2 similarly to RPE1 shows an upregulation by two mechanisms, one where the gene is upregulated in the hybrid model compared to the ecModel, and one where the Snf1 deletion shows a differential expression compared to the hybrid WT. The specific matrix of the gene regulatory network indicates that the mechanism is not related to a specific pathway. The regulation, of PFK2 and RPE1, is generated through a general regulatory effect caused by the network and not by regulation specifically acting on the gene itself.

Discussion

The effects of nutrient-induced signaling on metabolism play an important role in maintaining organismal homeostasis and consequently understanding human disease and aging. To gain a better understanding of the interconnectivity between nutrient signaling and metabolism, we have developed a hybrid model by combining a Boolean and an enzyme-constrained model of metabolism, using a regulatory network as a link. More specifically, we have implemented a Boolean signaling network that is responsive to glucose and nitrogen levels and an ecModel of yeast's central carbon metabolism. The proposed framework has been validated using available experimental data resulting in an increased predictive power on individual protein abundances in comparison to individual models alone. Further, we were able to characterize the cells' deviation from optimal protein allocation and flux distribution profiles. The model is capable of reproducing the regulatory effects that are associated with the Crabtree effect and glucose repression. In respiratory conditions, the model showed regulation of genes known to be differentially expressed in long-lived cells. This regulation was shown by the hybrid model to act via both Snf1 dependent and independent mechanisms. In addition, the model showed that during fermentation, enzyme utilization is the more important factor governing protein allocation, while in low glucose conditions robustness and control are prioritized.

The integration of regulatory constraints is resulting in a highly constrained hybrid model. The downside of this approach is connected to the lack of information regarding the regulatory effects of transcription factors activation. In this work we assume a uniform proportional action for all gene targets, together with the other constraints of the model, resulting in a rather low effect on the regulatory action. Despite this, the hybrid model shows improved predictive power for individual enzyme demands and can qualitatively reproduce regulatory effects associated with glucose repression in fermentation conditions, suggesting that with this framework we can gain novel insight into the interplay between signaling pathways and metabolism.

Another limitation is the inclusion of only the central carbon metabolism, a potential extension of this work would include the addition of other pathways responsive to glucose signaling, like glycerol metabolism and fatty acid synthesis, enabling also the study of the regulatory effect on these pathways specifically with relatively few modifications in the hybrid model.

The current state-of-the-art methods for absolute quantification of protein abundance typically yield high experimental errors, spanning even over orders of magnitude, when measuring external standards with proteins of known concentration [52–54]. Such measurement errors are comparable to the average error in prediction of individual enzyme levels by the hybrid model. Further comparison of enzyme usage profiles against proteomics datasets revealed that, incorporation of a regulatory layer over an ecModel induces a diversified isoenzymes utilization profile, supported by experimental evidence, in contrast to a purely optimality-based approach (pure ecModel) in which this is rarely observed, especially in non-protein limited conditions (cellular respiration at low dilution rates).

The hybrid model shows that under regulation the NADH to support the electron transport chain is partly coming from the cytosol with the help of the mitochondrial external NADH dehydrogenase, NDE2. Overexpression of NDI1, in contrast to NDE1, causes apoptosis-like cell death which can be repressed by growth on glucose-limited media [47]. In our model regulation acts on both NDE and NDI which will lower the need for NDI1 expression and thus causing apoptosis-like cell death. The hybrid model gives the ability to determine that the Snf1 pathway alone is responsible for the shift to the additional use of NDE and NDI instead of only NDI. Snf1 is active in glucose-limited media and thus would help mitigate the phenotype of overexpressed NDI1. With our approach, we can attribute this effect to the Snf1 pathway specifically which a metabolic model alone would not be able to predict. Further, connecting Snf1 with the respiration-restricted apoptotic activity described previously [47], hybrid model contributes to the understanding of the role of Snf1 in chronological aging [50]. Additionally, the hybrid model could also predict the additional use of ACS1, not predicted by the ecModel or the SNF1 deletion, by increasing the flux through the ACS reaction. This phenotype of Snf1 has been indicated as an important factor in caloric restriction related extension of chronological lifespan in yeast [50]. When comparing differentially expressed genes in cells with extended chronological life span with genes affected by regulation in the hybrid model, both genes differentially expressed in caloric restriction conditions were regulated by the Snf1 pathway in the hybrid model, further strengthening the Snf1 mediated mechanism of extended chronological lifespan after caloric restriction. RPE1 and PFK2 were found in two different conditions leading to extended chronological lifespan and also showed two mechanisms of regulation in the hybrid model through systems biology effects, one general and one mechanism working through the SNF1 pathway. Interestingly all caloric restriction related conditions show at least one mechanism of the regulation working via the SNF1 pathway. This exemplifies how we can confirm known and possibly predict novel connections between signaling and metabolism when combined in a coherent framework.

Futile fluxes in the cell have been examined previously within the constraints of osmotics, thermodynamics, and enzyme utilization [55], where the osmotics are putting a ceiling on the allowed metabolite concentrations in the cell while thermodynamics govern the net fluxes through reactions. The induced futile fluxes can be explained by the fact that regulation included in the hybrid model will force the cell to use some enzymes even above its pathway flux requirements, adding robustness of metabolism to a constantly changing environment. The increase in flux in both forward and backward directions (i.e the increased futile flux through reactions) implies that these enzymes are working closer to their equilibrium and thus have a low flux control over the pathway flux, while enzymes with a strong forward flux have large flux control [56]. This feature is also displayed by our hybrid model, in which all

enzymatic steps with induced futile fluxes exert null control over glucose uptake (FCCs = 0). More enzymes in a pathway working close to their equilibrium results in robustness against perturbations as well as allow the pathway to be controlled and regulated through a few enzymes, however, this happens at the expense of inefficient utilization of enzymes as the cell needs to spend more resources to sustain a pool of enzymes that are carrying both forward and backward fluxes [55, 57]. Our predictions of several glycolytic steps forced to operate closer to their equilibrium by regulation (high futile fluxes induced for TDH, PGK, and ENO in respiration, and TPI, PGK, and GPM in fermentation) agree with experimental studies on *E. coli*, iBMK cells and *Clostridia cellulyticum*, which have suggested the utility of near-equilibrium glycolytic steps not just for providing robustness to environmental changes but also for enhancing metabolic energy yield [58].

Computation of FCCs showed that in respiration the glucose flux is tightly dependent on the activity of the enzymatic steps in oxidative phosphorylation, a high-energy yield pathway. In contrast, in the fermentative condition flux control is split between PFK, PYK, PDC, and several steps in the TCA cycle. Interestingly, the FCCs in the TCA cycle are decreased by around half, after applying the regulatory constraints in the hybrid model, providing hints of the importance of enhancing robustness in this pathway at high growth rates due to increased demand for biomass precursors. The prevalence of the highest FCCs in fermentation for PFK, PYK, and PDC (for both the ecModel and the hybrid model) indicates their important role as modulators of flux balance between glycolysis, PPP, and fermentative pathways at highly demanding conditions, suggesting that when entering fermentation, the cell sacrifices robustness to favor efficient enzyme utilization.

Comparison of enzyme usage and flux distributions between models and across conditions reveals that the effects of regulation are generally stronger for the respiratory condition, causing the arisen of more and higher futile fluxes; turning on reaction steps that are not required by optimal metabolic allocation (purely ecModel) and inducing higher upregulation of fluxes. These findings suggest that metabolic phenotypes are majorly shaped by regulatory constraints in low glucose conditions, whilst enzymatic constraints play a major role when glucose is not the limiting resource.

It was also found that the regulatory layer diminishes the strong flux control that hexokinase isoforms have over glucose consumption in both low and high glucose conditions to 0. The hexokinases in yeast, specially HXK2, have a central role in glucose signaling. It works both as an effector in the Snf1 pathway and also actively participates in the repression complex together with Mig1 in glucose repression during high glucose conditions [59]. Intuitively, it would be practical if an enzyme having these central and diverse tasks in the cell would not have such a high FCC as can be seen with the ecModel. When small perturbations in enzyme activity or concentration have large effects on glucose consumptions, allocating this enzyme to other parts of the cell such as the nucleus, participating in the repression complex, would be energetically expensive. Given the central role of hexokinase in glucose signaling, this would be of interest for further investigation and future studies.

Overall, in this work, we have shown how the hybrid modeling framework integrating nutrient-sensing pathways and central carbon metabolism can not only improve individual model predictions but can also elucidate how single components in the dynamic signaling layer affect metabolism at steady-state. We tested our model against both respiring and fermenting conditions and could not only predict known phenomena but also find novel connections. This methodology can be used to connect both original and readily available models in yeast to look at the interactions between signaling and metabolism. This can be applied to genome-scale and on different subsystems of metabolism and for different signaling systems (e.g. macronutrients or osmotic stress sensing). The availability of genome-scale models for

different organisms is constantly growing and with our increasing understanding of signaling systems and regulatory networks, the methodology developed in the course of this work can be adapted to many other organisms. Hybrid models, like the one proposed here, also provide a framework for hypothesis testing, as we demonstrated by knocking out several components of the nutrient-induced signaling network. In summary, we developed a methodology to investigate intrinsically different systems, such as signaling and metabolism, integrated into the same model, gaining insight into how the interplay between them can have non-trivial effects.

Materials and methods

Boolean model of nutrient-induced signaling pathways

Based on an extensive literature review, a detailed topology of the nutrient-induced signaling pathways TORC1, SNF1 and PKA accounting also for their crosstalks was derived and formalized as a Boolean network model using a vector-based modelling approach [15] **TORC1**: [8, 60–74]; **SNF1:** [75–101]; **PKA:** [7, 8, 12, 38, 102–115]; **crosstalks**: [36, 38, 40, 109, 116, 117].

The model consists of four different components: metabolites, target genes, regulated enzymes, and proteins. For the regulated enzymes, presence and phosphorylation state were considered whereas metabolites and target genes were only described by a single binary value indicating their presence and transcriptional state respectively. The state vectors were translated into a single binary value indicating the components' activity, allowing a better graphical depiction. In total, the model comprises 5 metabolites, 10 groups of target genes, 6 enzymes whose activity is altered upon nutrient signaling, and 46 proteins belonging to PKA/cAMP, the SNF1, and the TORC1 pathway, for detailed description, see S1 Text and S1 Table.

The availability of glucose and nitrogen was used as an input to the model and is implemented as one vector of binary values for each nutrient. This input enables to simulate the induction of signaling under different nutrient conditions, for instance, the addition of glucose and nitrogen to starved cells is represented by the vector 0|1 for both nutrients respectively. Here, 0 represents the starved or low nutrient condition and 1 the nutrient-rich condition. Based on this input and the formulation of the Boolean rules, a cascade of state transitions is induced. The simulation was conducted using a synchronous updating scheme meaning that at each iteration, the state vectors are updated simultaneously. The algorithm stops if a Boolean steady state is reached at which no operation causes a change in the state vectors. This process is repeated for each pair of glucose and nitrogen availabilities whereby the reached steady state for each nutrient condition serves as an initial condition for the next nutrient condition.

Since for many of the included processes, no information on the mechanisms causing reversibility was available, especially a lack in knowledge on phosphatases reverting phosphorylation was observed [15], gap-filling was conducted by including else-statements. This ensures that a component's state vector changes again e.g. if the conditions causing its phosphorylation are not fulfilled anymore. This gap-filling process guarantees the functionality of the Boolean model in both directions, meaning the simulation of state transitions occurring when nutrients (glucose and nitrogen) are added to nutrient-depleted cells as well as when cells are starved for the respective nutrients. Crosstalk mechanisms between the pathways were formulated as if-statements and can be switched off (0) or on (1). Furthermore, a simulation of knockouts of the pathways' components is possible by setting the value indicating their presence to 0.

Enzyme-constrained metabolic model

A reduced stoichiometric model of *Saccharomyces cerevisiae's* central carbon and energy metabolism, including metabolites, reactions, genes, and their interactions accounting for

glycolysis, TCA cycle, oxidative phosphorylation, pentose phosphate, Leloir, and anaerobic excretion pathways, together with a representation of biomass formation, was taken as a network scaffold[29]. The metabolic model was further enhanced with enzyme constraints using the GECKO toolbox v1.3.5 [30], which considers enzymes as part of metabolic reactions, as they are occupied by metabolites for a given amount of time that is inversely proportional to the enzyme's turnover number (k_{cat}). Therefore, enzymes are incorporated as new "pseudo metabolites" and usage pseudo reactions are also introduced in order to represent their connection to a limited pool of protein mass available for metabolic enzymes. Moreover, all reversible reactions are split into two reactions with opposite directionalities in the ecModel, in order to account for the enzyme demands of backward fluxes. Several size metrics for the Boolean model, the metabolic network, and its enzyme-constrained version (ecModel) are shown in Table 1.

As the obtained ecModel has the same structure as any metabolic stoichiometric model, in which metabolites and reactions are connected by a stoichiometric matrix, the technique of flux balance analysis (FBA) can be used for quantitative prediction of intracellular reaction fluxes [118]. FBA assumes that the metabolic network operates on steady-state, i.e. no net accumulation of internal metabolites, due to the high turnover rate of metabolites when compared to cellular growth or environmental dynamics [119], therefore, by setting mass balances around each intracellular metabolite a homogenous system of linear equations is obtained. The second major assumption of FBA is that metabolic phenotypes are defined by underlying organizational principles, therefore an objective function is set as a linear combination of reaction fluxes which allows for obtaining a flux distribution by solving the following linear programming problem

$$\max: Z = C^T v$$

Subject to

$$S \cdot v = 0$$

$$lb \leq v \leq ub$$

Table 1. Size metrics for the Boolean, original metabolic model, and its enzyme-constrained version.

Boolean model	
Metabolites	5
Target gene groups	10
Enzyme PTMs	6
Proteins	46
Metabolic model	
Reactions	90
Metabolites	81
Genes	130
Cellular compartments	4
ecModel	
Reactions	324
Metabolites	111
Enzymes	127
Promiscuous enzymes	41
Reactions with isoenzymes	30
Enzyme complexes	11
Reactions w/Kcat	115

https://doi.org/10.1371/journal.pcbi.1008891.t001

Where C^T, is a transposed vector of integer coefficients for each flux in the objective function (Z); v, is the vector of reaction fluxes; S, is a stoichiometric matrix, representing metabolites as rows and reactions as columns; lb and ub are vectors of lower and upper bounds, respectively, for the reaction fluxes in the system. Additionally, the incorporation of enzyme constraints enables the connection between reaction fluxes and enzyme demands, which are constrained by the aforementioned pool of metabolic enzymes

$$v_i = \sum_j k_{cat_{ij}} \cdot e_j$$

$$\sum_j^p Mw_j \cdot e_j \leq f \cdot \sigma \cdot P_{tot}$$

Where $k_{cat_{ij}}$ is the turnover number of the enzyme j for the i-th reaction, as in some cases several enzymes can catalyze the same reaction (isoenzymes); e_j, is the usage rate for the enzyme j in mmol/gDw h^{-1}; Mw_j, represents the molecular weight of the enzyme j, in mmol/g; P_{tot}, is the total protein content in a yeast cell, corresponding to a value of 0.46 g_{prot}/gDw [120]; f, is the fraction of the total cell proteome that is accounted for in our ecModel, 0.1732 when using the integrated dataset for *S. cerevisiae* in paxDB as a reference [121]; and σ being an average saturation factor for all enzymes in the model.

This simple modeling formalism enables the incorporation of complex enzyme-reaction relations into the ecModel due to its matrix formulation, such as isoenzymes, which are different enzymes able to catalyze the same reaction; promiscuous enzymes, enzymes that can catalyze more than one reaction; and enzyme complexes, several enzyme subunits all needed to catalyze a given reaction.

ecModel curation

As the ecModel was generated by the automated pipeline of the GECKO toolbox, several of its components were curated to achieve predictions that are in agreement with experimental data at different dilution rates. Data on exchange reaction fluxes at increasing dilution rates, spanning both respiration and fermentative metabolic regimes [43] was used as a comparison basis. Additionally, all unused genes in the original metabolic network were removed and gene rules for lactose and galactose metabolism were corrected according to manually curated entries for *S. cerevisiae* available at the Swiss-Prot database [122]. Gene rules and metabolites stoichiometries (P/O ratio) in the oxidative phosphorylation pathway were also corrected according to the consensus genome-scale network reconstruction, Yeast8 [21].

The average saturation factor for the enzymes in the model was fitted to a value of 0.48, which allows the prediction of the experimental critical dilution rate (i.e. the onset of fermentative metabolism) at 0.285 h^{-1}. ATP requirements for biomass production were fitted by minimization of the median relative error in the prediction of exchange fluxes for glucose, oxygen, CO_2 and ethanol across dilution rates (0–0.4 h^{-1}), resulting in a linear relation depending on biomass formation from 18 to 25 mmol per gDw for respiratory conditions and from 25 to 30 mmol per gDw for the fermentative regime.

Hybrid model

A hybrid model consists of the Boolean model connected with the ecModel through a transcriptional layer that regulates its constraints on protein allocation (Fig 1). The active transcription factors act on the upper or lower bounds of the enzyme usage pseudo reaction depending on down- or up- regulation, respectively. The magnitude of the induced

perturbations is calculated according to previously calculated enzyme usage variability ranges, subject to a given growth rate and optimal glucose rate, expressed as

Upregulation:

$$lb_{e_i}^{reg} = e_i^{opt} + RF * (e_i^{max} - e_i^{min})$$

Downregulation:

$$ub_{e_i}^{reg} = e_i^{opt} - RF * (e_i^{max} - e_i^{min})$$

Where $lb_{e_i}^{reg}$ and $ub_{e_i}^{reg}$ represent the lower and upper bounds for the usage pseudo reaction of enzyme *i* in the regulated model; e_i^{opt}, is a parsimonious usage for enzyme *i* for a given growth and glucose uptake rates; *RF*, corresponds to a regulation factor between 0 and 1; e_i^{max} and e_i^{min} are the maximum and minimum allowable usages for enzyme *i* under the specified conditions.

A distribution of parsimonious enzyme usages is obtained by applying the rationale of the parsimonious FBA technique [123], which explicitly minimizes the total protein burden that sustains a given metabolic state (i.e. fixed growth and nutrient uptake rates).

To connect the transcription factor activity with gene regulation we extracted regulation information from YEASTRACT and set a regulation level of 5% of the enzyme usage variability range for the simulations. When several transcription factors affect the same gene, the effects are summed up and the resulting sum is used as a basis for constraint. For example, if a gene is downregulated by two transcription factors (-2) and upregulated by one transcription factor (+1), the net sum would be (-1), thus the gene will be downregulated. In our model, an absolute sum higher than 1 will not cause a stronger regulation, as this additive process is just implemented to define the directionality of a regulatory effect.

2.5 Enzyme usage variability analysis

As metabolic networks are highly redundant and interconnected, the use of purely stoichiometric constraints usually leads to an underdetermined system with infinite solutions [124], in a typical FBA problem it is common that even for an optimal value of the objective function, several reactions in the network can take any value within a "feasible" range, such ranges can be explored by flux variability analysis [24].

In this study, enzyme usage variability ranges for all of the individual enzymes are calculated by fixing a minimal glucose uptake flux, for a given fixed dilution rate, and then running sequential maximization and minimization for each enzyme usage pseudo reaction.

$$enzyme\ usage\ variability\ range = e_i^{max} - e_i^{min}$$

Subject to:

$$v_{Glc_{IN}} = lb_{Glc_{IN}} = ub_{Glc_{IN}} = v_{Glc_{IN}}$$

$$v_{bio} = lb_{bio} = ub_{bio} = D_{rate}$$

This approach allows the identification of enzymes that are either tightly constrained, highly variable, or even not usable at optimal levels of biomass yield.

Simulations

Cellular growth on chemostat conditions using minimal media with glucose as a carbon source, at varying dilution rates from 0 to 0.4 h^{-1}, was simulated with the multiscale model by the following sequence of steps:

1. Initially, the desired dilution rate is set as both lower and upper bounds for the growth pseudo reaction and the glucose uptake rate is minimized, assuming that cells maximize biomass production yield when glucose is limited [125, 126]

$$\min : \ v_{Glc_{IN}}$$

   Subject to

$$D_{rate} \leq v_{bio} \leq D_{rate}$$

2. The obtained optimal uptake rate ($v^{min}_{Glc_{IN}}$) is then used as a basis to estimate a range of uptake flux to further constrain the ecModel.

$$v^{min}_{Glc_{IN}} \leq v_{Glc_{IN}} \leq (1 + SF) * v^{min}_{Glc_{IN}}$$

   As $v^{min}_{Glc_{IN}}$ represents the minimum uptake rate allowed by the stoichiometric and enzymatic constraints of the metabolic network, possible deviations from optimal behavior may be induced by regulatory circuits. To allow the Boolean model to reallocate enzyme levels a suboptimality factor (*SF*) of 15% was used to set an upper bound for $v_{Glc_{IN}}$.

3. The ecModel is connected to the glucose-sensing Boolean model through the glucose uptake rate. At the critical dilution rate, the glucose uptake rate obtained by the ecModel is 3.2914 mmol/gDw h, this value is used as a threshold to define a "low" or "high" glucose level input in the Boolean model, represented as 0 and 1, respectively. For each dilution rate, the initial value of $v^{min}_{Glc_{IN}}$ is calculated and fed to the regulatory network, which runs a series of synchronous update steps until a steady-state is reached.

4. At steady state, the regulatory network indicates the enzyme usages that should be up and downregulated, for which new usage bounds are set as described above.

5. A final FBA simulation is run by minimizing the glucose uptake rate, subject to a fixed dilution rate, and the newly regulated enzyme usage bounds.

Gene deletions can also be set in the Boolean module and will result in activation or inactivation of transcription factors which then affect the constraints on the FBA model. We ran four simulations of deletion strains as follows: TOR1 and TOR2 (TOR deletion), Snf1 (SNF1 deletion), Tpk1, Tpk2, and Tpk3 (PKA deletion), and Reg1(Reg1 deletion).

Proteomics analysis

Protein abundance data on respiratory and fermentative conditions were compared to protein usage predictions by the hybrid model to assess its performance. For the respiration phase, absolute protein abundances were taken from a study of yeast growing under glucose-limited chemostat conditions at 30˚C on minimal mineral medium with a dilution rate of 0.1 h^{-1} [44].

For the fermentation phase, a proteomics dataset was taken from a batch culture using minimal media with 2% glucose and harvested at an optical density (OD) of 0.6 [45]. The dataset

given as relative abundances was then rescaled to relative protein abundances in the whole-cell according to integrated data available for *S. cerevisiae* in PaxDB [127], and finally converted to absolute units of mmol/gDw using the "total protein approach" [128].

We used three metrics for comparing the simulations with the proteomics data, the Pearson correlation coefficient (PCC), two-sample Kolmogorov-Smirnov (KS) test, and the mean of the absolute $\log_{10}$-transformed ratios between predicted and measured values (r). The PCC and the significance of the PCC were determined by a permutation test of n = 2000. The pathway enrichments were done using YeastMine [129] with the Holm-Bonferroni test correction and a max p-value of 0.05.

Flux control coefficients

To investigate the relationship between enzyme activities and a given metabolic flux, control coefficients can be calculated for each enzyme in the model according to the definition given by metabolic control analysis (MCA) [56]:

$$FCC_{ij} = \frac{a_i}{v_j}\frac{\partial v_j}{\partial a_i}$$

In which $a_i = k_{cat_{ij}} e_i$ represents the activity of the *i-th* enzyme and v_j is the flux carried by the *j-th* reaction. These coefficients represent the sensitivity of a given metabolic flux to perturbations on enzyme activities, providing a quantitative measure on the control that each enzyme exerts on specific fluxes.

As ecModels include enzyme activities explicitly in their structure, flux control coefficients can be approximated by inducing small perturbations on individual enzyme usages:

$$FCC_{ij} \approx \frac{k_{cat_{ij}} e_i}{v_j}\frac{\Delta v_j}{\Delta(k_{cat_{ij}} e_i)}$$

In our hybrid model, perturbations on individual enzyme usages (e_i) are induced in relation to a parsimonious usage (e_i^*) which is compatible with a given set of constraints

$$FCC_{ij} = \frac{e_i^*}{v_j^*}\frac{\Delta v_j}{\Delta(e_i - e_i^*)}$$

Perturbations equivalent to 0.1% of the parsimonious usage are used for each enzyme. For those cases in which the previously applied constraints do not allow such modification in a given enzyme usage, their activity is then perturbed by operating on the corresponding turnover number for the enzyme-reaction pair ($k^*_{cat_{ij}} = 0.001 * k_{cat_{ij}}$) to simulate a perturbation in their overall activity.

Supporting information

S1 Text. Supporting information on the Boolean layer. Includes a detailed description of mechanisms reflected in the Boolean model of nutrient signaling as well as open questions of dynamics and model gaps.
(DOCX)

S2 Text. Supporting Information on the hybrid model. Includes detailed information on the analysis of protein prediction and deletion strain simulations.
(DOCX)

S1 Fig. Transition map of components in the Boolean model separated by their respective pathway where the blue color indicates activity. Simulations are made with all crosstalk turned on. Panel (A) shows the simulation dynamics going from nutrient-depleted conditions to nutrient-rich conditions and panel (B) shows the simulation dynamics going from nutrient-rich conditions to nutrient depletion.
(PDF)

S2 Fig. Steady-state map of components in the Boolean model when knock-out (KO) strains are simulated from wild type (WT) to KO where the blue color indicates activity. Panel (A) shows the KO behavior in low nutrient conditions compared to the WT and panel (B) show the KO behavior in high nutrient conditions compared to the WT
(PDF)

S3 Fig. Exchange fluxes for the hybrid model plotted over experimental data. Simulations showed a median relative error of 9.82% in the whole range of dilution rates from 0 to 0.4 h-1.
(PDF)

S4 Fig. The fluxes through the core reactions in the metabolism are represented by the width of the connectors where dotted lines represent zero flux. The color of the connectors represents the change in flux from the wild type (WT) hybrid model compared to the SNF1 deletion hybrid model. The FCCs are represented in the model where the WT is compared to the SNF1 deletion case.
(PDF)

S1 Table. Rules and references associated with any field of any of the Boolean vectors in the Boolean module.
(DOCX)

S2 Table. Summary of the statistics done comparing the ecModel and the hybrid model in their ability to predict protein abundance.
(DOCX)

S1 Data. Data related to enzyme usages and protein prediction.
(XLSX)

S2 Data. Data related to fluxes.
(XLSX)

S3 Data. Data related to FCC.
(XLSX)

S4 Data. Data related to mutant strain simulations.
(XLSX)

Acknowledgments

We would like to thank members of the Hohmann, Nielsen, and Cvijovic labs for valuable input. Special thanks to Avlant Nilsson for his contributions to the curation of the original metabolic network used in this study and valuable discussions on the role of enzyme constraints.

Author Contributions

Conceptualization: Linnea Österberg, Marija Cvijovic.

Data curation: Linnea Österberg, Iván Domenzain, Julia Münch.

Formal analysis: Linnea Österberg, Iván Domenzain.

Funding acquisition: Jens Nielsen, Stefan Hohmann, Marija Cvijovic.

Investigation: Linnea Österberg, Iván Domenzain, Julia Münch.

Methodology: Linnea Österberg, Iván Domenzain, Julia Münch.

Project administration: Linnea Österberg, Marija Cvijovic.

Resources: Jens Nielsen, Marija Cvijovic.

Software: Linnea Österberg, Iván Domenzain.

Supervision: Jens Nielsen, Stefan Hohmann, Marija Cvijovic.

Validation: Linnea Österberg, Iván Domenzain.

Visualization: Linnea Österberg, Iván Domenzain, Julia Münch.

Writing – original draft: Linnea Österberg, Iván Domenzain.

Writing – review & editing: Jens Nielsen, Stefan Hohmann, Marija Cvijovic.

References

1. Walpole J, Papin JA, Peirce SM. Multiscale Computational Models of Complex Biological Systems. Annu Rev Biomed Eng. 2013 Jul 11; 15(1):137–54. https://doi.org/10.1146/annurev-bioeng-071811-150104 PMID: 23642247
2. Wang YP, Lei QY. Metabolite sensing and signaling in cell metabolism. Vol. 3, Signal Transduction and Targeted Therapy. Springer Nature; 2018. p. 1–9. https://doi.org/10.1038/s41392-017-0001-6 PMID: 29527327
3. Coughlan KA, Valentine RJ, Ruderman NB, Saha AK. AMPK activation: A therapeutic target for type 2 diabetes? Vol. 7, Diabetes, Metabolic Syndrome and Obesity: Targets and Therapy. Dove Press; 2014. p. 241–53.
4. Li W, Saud SM, Young MR, Chen G, Hua B. Targeting AMPK for cancer prevention and treatment. Oncotarget. 2015 Apr; 6(10):7365–78. https://doi.org/10.18632/oncotarget.3629 PMID: 25812084
5. Salminen A, Kaarniranta K. AMP-activated protein kinase (AMPK) controls the aging process via an integrated signaling network. Vol. 11, Ageing Research Reviews. Elsevier; 2012. p. 230–41. https://doi.org/10.1016/j.arr.2011.12.005 PMID: 22186033
6. Steinberg GR, Kemp BE. AMPK in Health and Disease. Physiol Rev. 2009; 89(3):1025–78. https://doi.org/10.1152/physrev.00011.2008 PMID: 19584320
7. Conrad M, Schothorst J, Kankipati HN, Van Zeebroeck G, Rubio-Texeira M, Thevelein JM. Nutrient sensing and signaling in the yeast Saccharomyces cerevisiae. Vol. 38, FEMS Microbiology Reviews. Wiley-Blackwell; 2014. p. 254–99. https://doi.org/10.1111/1574-6976.12065 PMID: 24483210
8. Broach JR. Nutritional control of growth and development in yeast. Vol. 192, Genetics. Genetics Society of America; 2012. p. 73–105. https://doi.org/10.1534/genetics.111.135731 PMID: 22964838
9. Ashrafi K, Lin SS, Manchester JK, Gordon JI. Sip2p and its partner Snf1p kinase affect aging in S. cerevisiae. Genes Dev. 2000; 14(15):1872–85. PMID: 10921902
10. Hedbacker K, Carlson M. SNF1/AMPK pathways in yeast. Vol. 13, Frontiers in Bioscience. NIH Public Access; 2008. p. 2408–20. https://doi.org/10.2741/2854 PMID: 17981722
11. Lempiäinen H, Uotila A, Urban J, Dohnal I, Ammerer G, Loewith R, et al. Sfp1 Interaction with TORC1 and Mrs6 Reveals Feedback Regulation on TOR Signaling. Mol Cell. 2009 Mar 27; 33(6):704–16. https://doi.org/10.1016/j.molcel.2009.01.034 PMID: 19328065
12. Marion RM, Regev A, Segal E, Barash Y, Koller D, Friedman N, et al. Sfp1 is a stress- and nutrient-sensitive regulator of ribosomal protein gene expression. Proc Natl Acad Sci U S A. 2004 Oct 5; 101 (40):14315–22. https://doi.org/10.1073/pnas.0405353101 PMID: 15353587
13. Welkenhuysen N, Borgqvist J, Backman M, Bendrioua L, Goksör M, Adiels CB, et al. Single-cell study links metabolism with nutrient signaling and reveals sources of variability. BMC Syst Biol. 2017;11(1). https://doi.org/10.1186/s12918-017-0392-6 PMID: 28122551

14. Shashkova S, Welkenhuysen N, Hohmann S. Molecular communication: crosstalk between the Snf1 and other signaling pathways. FEMS Yeast Res. 2015;15. https://doi.org/10.1093/femsyr/fov026 PMID: 25994786
15. Welkenhuysen N, Schnitzer B, Österberg L, Cvijovic M. Robustness of nutrient signaling is maintained by interconnectivity between signal transduction pathways. Front Physiol. 2019 Jan 21; 10 (JAN):1964. https://doi.org/10.3389/fphys.2018.01964 PMID: 30719010
16. Nielsen J. Systems Biology of Metabolism. Annu Rev Biochem. 2017;
17. Christensen TS, Oliveira AP, Nielsen J. Reconstruction and logical modeling of glucose repression signaling pathways in Saccharomyces cerevisiae. BMC Syst Biol. 2009 Jan 14; 3:7. https://doi.org/10.1186/1752-0509-3-7 PMID: 19144179
18. Sulaimanov N, Klose M, Busch H, Boerries M. Understanding the mTOR signaling pathway via mathematical modeling. Vol. 9, Wiley Interdisciplinary Reviews: Systems Biology and Medicine. Wiley-Blackwell; 2017. https://doi.org/10.1002/wsbm.1379 PMID: 28186392
19. Siegle L, Schwab JD, Kühlwein SD, Lausser L, Tümpel S, Pfister AS, et al. A Boolean network of the crosstalk between IGF and Wnt signaling in aging satellite cells. PLoS One. 2018 Mar 1; 13(3). https://doi.org/10.1371/journal.pone.0195126 PMID: 29596489
20. Romers J, Thieme S, Münzner U, Krantz M. A scalable method for parameter-free simulation and validation of mechanistic cellular signal transduction network models. npj Syst Biol Appl. 2020 Dec 1; 6(1). https://doi.org/10.1038/s41540-019-0120-5 PMID: 31934349
21. Lu H, Li F, Sánchez BJ, Zhu Z, Li G, Domenzain I, et al. A consensus S. cerevisiae metabolic model Yeast8 and its ecosystem for comprehensively probing cellular metabolism. Nat Commun. 2019 Dec 1; 10(1):1–13. https://doi.org/10.1038/s41467-018-07882-8 PMID: 30602773
22. Yilmaz LS, Walhout AJ. Metabolic network modeling with model organisms. Vol. 36, Current Opinion in Chemical Biology. Elsevier Ltd; 2017. p. 32–9. https://doi.org/10.1016/j.cbpa.2016.12.025 PMID: 28088694
23. Monk JM, Lloyd CJ, Brunk E, Mih N, Sastry A, King Z, et al. iML1515, a knowledgebase that computes Escherichia coli traits. Vol. 35, Nature Biotechnology. 2017. p. 904–8. https://doi.org/10.1038/nbt.3956 PMID: 29020004
24. Mahadevan R, Schilling CH. The effects of alternate optimal solutions in constraint-based genome-scale metabolic models. Metab Eng. 2003; 5(4):264–76. https://doi.org/10.1016/j.ymben.2003.09.002 PMID: 14642354
25. Adadi R, Volkmer B, Milo R, Heinemann M, Shlomi T. Prediction of microbial growth rate versus biomass yield by a metabolic network with kinetic parameters. PLoS Comput Biol. 2012; 8(7). https://doi.org/10.1371/journal.pcbi.1002575 PMID: 22792053
26. Beg QK, Vazquez A, Ernst J, de Menezes MA, Bar-Joseph Z, Barabási AL, et al. Intracellular crowding defines the mode and sequence of substrate uptake by Escherichia coli and constrains its metabolic activity. Proc Natl Acad Sci U S A. 2007; 104(31):12663–8. https://doi.org/10.1073/pnas.0609845104 PMID: 17652176
27. Bekiaris PS, Klamt S. Automatic construction of metabolic models with enzyme constraints. BMC Bioinformatics. 2020; https://doi.org/10.1186/s12859-019-3329-9 PMID: 31937255
28. Massaiu I, Pasotti L, Sonnenschein N, Rama E, Cavaletti M, Magni P, et al. Integration of enzymatic data in Bacillus subtilis genome-scale metabolic model improves phenotype predictions and enables in silico design of poly-γ-glutamic acid production strains. Microb Cell Fact. 2019; https://doi.org/10.1186/s12934-018-1052-2 PMID: 30626384
29. Nilsson A, Nielsen J. Metabolic Trade-offs in Yeast are Caused by F1F0-ATP synthase. Sci Rep. 2016; 6:1–11. https://doi.org/10.1038/s41598-016-0001-8 PMID: 28442746
30. Sánchez BJ, Zhang C, Nilsson A, Lahtvee P, Kerkhoven EJ, Nielsen J. Improving the phenotype predictions of a yeast genome-scale metabolic model by incorporating enzymatic constraints. Mol Syst Biol. 2017; 13(8):935. https://doi.org/10.15252/msb.20167411 PMID: 28779005
31. Shlomi T, Eisenberg Y, Sharan R, Ruppin E. A genome-scale computational study of the interplay between transcriptional regulation and metabolism. Mol Syst Biol. 2007; 3:101. https://doi.org/10.1038/msb4100141 PMID: 17437026
32. Banos DT, Trébulle P, Elati M. Integrating transcriptional activity in genome-scale models of metabolism. BMC Syst Biol. 2017 Dec 21; 11(Suppl 7).
33. Marmiesse L, Peyraud R, Cottret L. FlexFlux: combining metabolic flux and regulatory network analyses. BMC Syst Biol. 2015 Dec 15; 9(1):93. https://doi.org/10.1186/s12918-015-0238-z PMID: 26666757

34. Bose T, Das C, Dutta A, Mahamkali V, Sadhu S, Mande SS. Understanding the role of interactions between host and Mycobacterium tuberculosis under hypoxic condition: an in silico approach. BMC Genomics. 2018 Dec 27; 19(1):555. https://doi.org/10.1186/s12864-018-4947-8 PMID: 30053801

35. Pedruzzi I, Dubouloz F, Cameroni E, Wanke V, Roosen J, Winderickx J, et al. TOR and PKA Signaling Pathways Converge on the Protein Kinase Rim15 to Control Entry into G0. Mol Cell. 2003; 12 (6):1607–13. https://doi.org/10.1016/s1097-2765(03)00485-4 PMID: 14690612

36. Cherry JR, Johnson TR, Dollard C, Shuster JR, Denis CL. Cyclic AMP-dependent protein kinase phosphorylates and inactivates the yeast transcriptional activator ADR1. Cell. 1989 Feb 10; 56(3):409–19. https://doi.org/10.1016/0092-8674(89)90244-4 PMID: 2644045

37. Thompson-Jaeger S, Francois J, Gaughran JP, Tatchell K. Deletion of SNF1 affects the nutrient response of yeast and resembles mutations which activate the adenylate cyclase pathway. Genetics. 1991; 129(3):697–706. PMID: 1752415

38. Hughes Hallett JE, Luo X, Capaldi AP. State transitions in the TORC1 signaling pathway and information processing in Saccharomyces cerevisiae. Genetics. 2014 Oct 1; 198(2):773–86. https://doi.org/10.1534/genetics.114.168369 PMID: 25085507

39. Dombek KM, Voronkova V, Raney A, Young ET. Functional analysis of the yeast Glc7-binding protein Reg1 identifies a protein phosphatase type 1-binding motif as essential for repression of ADH2 expression. Mol Cell Biol. 1999 Sep; 19(9):6029–40. https://doi.org/10.1128/mcb.19.9.6029 PMID: 10454550

40. Barrett L, Orlova M, Maziarz M, Kuchin S. Protein kinase a contributes to the negative control of SNF1 protein kinase in saccharomyces cerevisiae. Eukaryot Cell. 2012 Feb; 11(2):119–28. https://doi.org/10.1128/EC.05061-11 PMID: 22140226

41. Robertson LS, Fink GR. The three yeast A kinases have specific signaling functions in pseudohyphal growth. Proc Natl Acad Sci U S A. 1998 Nov 10; 95(23):13783–7. https://doi.org/10.1073/pnas.95.23.13783 PMID: 9811878

42. Barbet NC, Schneider U, Helliwell SB, Stansfield I, Tuite MF, Hall MN. TOR controls translation initiation and early G1 progression in yeast. Mol Biol Cell. 1996; 7(1):25–42. https://doi.org/10.1091/mbc.7.1.25 PMID: 8741837

43. Van Hoek P, Van Dijken JP, Pronk JT. Effect of specific growth rate on fermentative capacity of baker's yeast. Appl Environ Microbiol. 1998 Nov; 64(11):4226–33. https://doi.org/10.1128/AEM.64.11.4226-4233.1998 PMID: 9797269

44. Doughty TW, Domenzain I, Millan-Oropeza A, Montini N, de Groot PA, Pereira R, et al. Stress-induced expression is enriched for evolutionarily young genes in diverse budding yeasts. Nat Commun. 2020;

45. Paulo JA, O'Connell JD, Everley RA, O'Brien J, Gygi MA, Gygi SP. Quantitative mass spectrometry-based multiplexing compares the abundance of 5000 S. cerevisiae proteins across 10 carbon sources. J Proteomics. 2016 Oct 4; 148:85–93. https://doi.org/10.1016/j.jprot.2016.07.005 PMID: 27432472

46. Meilă M. Comparing clusterings-an information based distance. J Multivar Anal. 2007;

47. Li W, Sun L, Liang Q, Wang J, Mo W, Zhou B. Yeast AMID homologue Ndi1p displays respiration-restricted apoptotic activity and is involved in chronological aging. Mol Biol Cell. 2006 Apr 25; 17 (4):1802–11. https://doi.org/10.1091/mbc.e05-04-0333 PMID: 16436509

48. de Alteriis E, Cartenì F, Parascandola P, Serpa J, Mazzoleni S. Revisiting the Crabtree/Warburg effect in a dynamic perspective: a fitness advantage against sugar-induced cell death. Vol. 17, Cell Cycle. 2018. p. 688–701. https://doi.org/10.1080/15384101.2018.1442622 PMID: 29509056

49. Tanner LB, Goglia AG, Wei MH, Sehgal T, Parsons LR, Park JO, et al. Four Key Steps Control Glycolytic Flux in Mammalian Cells. Cell Syst. 2018; https://doi.org/10.1016/j.cels.2018.06.003 PMID: 29960885

50. Wierman MB, Maqani N, Strickler E, Li M, Smith JS. Caloric Restriction Extends Yeast Chronological Life Span by Optimizing the Snf1 (AMPK) Signaling Pathway. Mol Cell Biol. 2017 Jul 1; 37(13). https://doi.org/10.1128/MCB.00562-16 PMID: 28373292

51. Wierman MB, Matecic M, Valsakumar V, Li M, Smith DL, Bekiranov S, et al. Functional genomic analysis reveals overlapping and distinct features of chronologically long-lived yeast populations. Aging (Albany NY). 2015; 7(3):177–94. https://doi.org/10.18632/aging.100729 PMID: 25769345

52. Arike L, Valgepea K, Peil L, Nahku R, Adamberg K, Vilu R. Comparison and applications of label-free absolute proteome quantification methods on Escherichia coli. J Proteomics. 2012; https://doi.org/10.1016/j.jprot.2012.06.020 PMID: 22771841

53. Ravikrishnan A, Raman K. Critical assessment of genome-scale metabolic networks: The need for a unified standard. Brief Bioinform. 2015; https://doi.org/10.1093/bib/bbv003 PMID: 25725218

54. Sánchez BJ, Lahtvee P-J, Campbell K, Kasvandik S, Yu R, Domenzain I, et al. Benchmarking accuracy and precision of intensity-based absolute quantification of protein abundances in

Saccharomyces cerevisiae bioRxiv. 2020 Jan;2020.03.23.998237. https://doi.org/10.1002/pmic.202000093 PMID: 33452728

55. Park JO, Rubin SA, Xu YF, Amador-Noguez D, Fan J, Shlomi T, et al. Metabolite concentrations, fluxes and free energies imply efficient enzyme usage. Nat Chem Biol. 2016 Jul 1; 12(7):482–9. https://doi.org/10.1038/nchembio.2077 PMID: 27159581

56. Kacser H, Burns JA, Fell DA. The control of flux. In: Biochemical Society Transactions. 1995. https://doi.org/10.1042/bst0230341 PMID: 7672373

57. Noor E, Flamholz A, Bar-Even A, Davidi D, Milo R, Liebermeister W. The Protein Cost of Metabolic Fluxes: Prediction from Enzymatic Rate Laws and Cost Minimization. PLoS Comput Biol. 2016; https://doi.org/10.1371/journal.pcbi.1005167 PMID: 27812109

58. Park JO, Tanner LB, Wei MH, Khana DB, Jacobson TB, Zhang Z, et al. Near-equilibrium glycolysis supports metabolic homeostasis and energy yield. Nat Chem Biol. 2019; https://doi.org/10.1038/s41589-019-0364-9 PMID: 31548693

59. Vega M, Riera A, Fernández-Cid A, Herrero P, Moreno F. Hexokinase 2 Is an intracellular glucose sensor of yeast cells that maintains the structure and activity of mig1 protein repressor complex. J Biol Chem. 2016; https://doi.org/10.1074/jbc.M115.711408 PMID: 26865637

60. Sanz P, Alms GR, Haystead TAJ, Carlson M. Regulatory Interactions between the Reg1-Glc7 Protein Phosphatase and the Snf1 Protein Kinase. Mol Cell Biol. 2000 Feb 15; 20(4):1321–8. https://doi.org/10.1128/mcb.20.4.1321-1328.2000 PMID: 10648618

61. Ludin K, Jiang R, Carlson M. Glucose-regulated interaction of a regulatory subunit of protein phosphatase 1 with the Snf1 protein kinase in Saccharomyces cerevisiae. Proc Natl Acad Sci U S A. 1998 May 26; 95(11):6245–50. https://doi.org/10.1073/pnas.95.11.6245 PMID: 9600950

62. Hong SP, Leiper FC, Woods A, Carling D, Carlson M. Activation of yeast Snf1 and mammalian AMP-activated protein kinase by upstream kinases. Proc Natl Acad Sci U S A. 2003 Jul 22; 100(15):8839–43. https://doi.org/10.1073/pnas.1533136100 PMID: 12847291

63. Sutherland CM, Hawley SA, McCartney RR, Leech A, Stark MJR, Schmidt MC, et al. Elm1p is one of three upstream kinases for the Saccharomyces cerevisiae SNF1 complex. Curr Biol. 2003 Aug 5; 13 (15):1299–305. https://doi.org/10.1016/s0960-9822(03)00459-7 PMID: 12906789

64. Leverentz MK, Reece RJ. Phosphorylation of Zn(II)2Cys6 proteins: A cause or effect of transcriptional activation? Biochem Soc Trans. 2006 Nov; 34(5):794–7. https://doi.org/10.1042/BST0340794 PMID: 17052200

65. Turcotte B, Liang XB, Robert F, Soontorngun N. Transcriptional regulation of nonfermentable carbon utilization in budding yeast. Vol. 10, FEMS Yeast Research. PMC Canada manuscript submission; 2010. p. 2–13.

66. MacPherson S, Larochelle M, Turcotte B. A Fungal Family of Transcriptional Regulators: the Zinc Cluster Proteins. Microbiol Mol Biol Rev. 2006 Sep 1; 70(3):583–604. https://doi.org/10.1128/MMBR.00015-06 PMID: 16959962

67. Westholm JO, Nordberg N, Murén E, Ameur A, Komorowski J, Ronne H. Combinatorial control of gene expression by the three yeast repressors Mig1, Mig2 and Mig3. BMC Genomics. 2008 Dec 16; 9 (SUPPL. 2):601. https://doi.org/10.1186/1471-2164-9-601 PMID: 19087243

68. Santangelo GM. Glucose Signaling in Saccharomyces cerevisiae. Microbiol Mol Biol Rev. 2006 Mar 1; 70(1):253–82. https://doi.org/10.1128/MMBR.70.1.253-282.2006 PMID: 16524925

69. Schüller HJ. Transcriptional control of nonfermentative metabolism in the yeast Saccharomyces cerevisiae. Curr Genet. 2003 Jun 1; 43(3):139–60. https://doi.org/10.1007/s00294-003-0381-8 PMID: 12715202

70. Soontorngun N, Baramee S, Tangsombatvichit C, Thepnok P, Cheevadhanarak S, Robert F, et al. Genome-wide location analysis reveals an important overlap between the targets of the yeast transcriptional regulators Rds2 and Adr1. Biochem Biophys Res Commun. 2012 Jul 13; 423(4):632–7. https://doi.org/10.1016/j.bbrc.2012.05.151 PMID: 22687600

71. Kacherovsky N, Tachibana C, Amos E, Fox D, Young ET. Promoter binding by the Adr1 transcriptional activator may be regulated by phosphorylation in the DNA-binding region. PLoS One. 2008 Sep 15; 3 (9). https://doi.org/10.1371/journal.pone.0003213 PMID: 18791642

72. Smith JJ, Miller LR, Kreisberg R, Vazquez L, Wan Y, Aitchison JD. Environment-responsive transcription factors bind subtelomeric elements and regulate gene silencing. Mol Syst Biol. 2011; 7:455. https://doi.org/10.1038/msb.2010.110 PMID: 21206489

73. Fernández-García P, Peláez R, Herrero P, Moreno F. Phosphorylation of Yeast Hexokinase 2 Regulates Its Nucleocytoplasmic Shuttling *. J Biol Chem. 2012;

74. Woods A, Munday MR, Scott J, Yang X, Carlson M, Carling D. Yeast SNF1 is functionally related to mammalian AMP-activated protein kinase and regulates acetyl-CoA carboxylase in vivo. J Biol Chem. 1994 Jul 29; 269(30):19509–15. PMID: 7913470

75. Jones S, Vignais ML, Broach JR. The CDC25 protein of Saccharomyces cerevisiae promotes exchange of guanine nucleotides bound to ras. Mol Cell Biol. 1991 May 1; 11(5):2641–6. https://doi.org/10.1128/mcb.11.5.2641 PMID: 2017169

76. Robinson LC, Gibbs JB, Marshall MS, Sigal IS, Tatchell K. CDC25: A component of the RAS-adenylate cyclase pathway in Saccharomyces cerevisiae. Science (80-). 1987 Mar 6; 235(4793):1218–21. https://doi.org/10.1126/science.3547648 PMID: 3547648

77. Broek D, Toda T, Michaeli T, Levin L, Birchmeier C, Zoller M, et al. The S. cerevisiae CDC25 gene product regulates the RAS/adenylate cyclase pathway. Cell. 1987 Mar 13; 48(5):789–99. https://doi.org/10.1016/0092-8674(87)90076-6 PMID: 3545497

78. Tanaka K, Nakafuku M, Tamanoi F, Kaziro Y, Matsumoto K, Toh-e A. IRA2, a second gene of Saccharomyces cerevisiae that encodes a protein with a domain homologous to mammalian ras GTPase-activating protein. Mol Cell Biol. 1990 Aug 1; 10(8):4303–13. https://doi.org/10.1128/mcb.10.8.4303 PMID: 2164637

79. Tanaka K, Nakafuku M, Satoh T, Marshall MS, Gibbs JB, Matsumoto K, et al. S. cerevisiae genes IRA1 and IRA2 encode proteins that may be functionally equivalent to mammalian ras GTPase activating protein. Cell. 1990 Mar 9; 60(5):803–7. https://doi.org/10.1016/0092-8674(90)90094-u PMID: 2178777

80. Tanaka K, Matsumoto K, Toh-E A. IRA1, an inhibitory regulator of the RAS-cyclic AMP pathway in Saccharomyces cerevisiae. Mol Cell Biol. 1989 Feb 1; 9(2):757–68. https://doi.org/10.1128/mcb.9.2.757 PMID: 2540426

81. Kraakman L, Lemaire K, Ma P, Teunlssen AWRH, Donaton MCV, Van Dijck P, et al. A Saccharomyces cerevisiae G-protein coupled receptor, Gpr1, is specifically required for glucose activation of the cAMP pathway during the transition to growth on glucose. Mol Microbiol. 1999; 32(5):1002–12. https://doi.org/10.1046/j.1365-2958.1999.01413.x PMID: 10361302

82. Colombo S, Ma P, Cauwenberg L, Winderickx J, Crauwels M, Teunissen A, et al. Involvement of distinct G-proteins, Gpa2 and Ras, in glucose- and intracellular acidification-induced cAMP signalling in the yeast Saccharomyces cerevisiae. EMBO J. 1998 Jun 15; 17(12):3326–41. https://doi.org/10.1093/emboj/17.12.3326 PMID: 9628870

83. Kataoka T, Broek D, Wigler M. DNA sequence and characterization of the S. cerevisiae gene encoding adenylate cyclase. Cell. 1985 Dec 1; 43(2 PART 1):493–505. https://doi.org/10.1016/0092-8674(85)90179-5 PMID: 2934138

84. Toda T, Uno I, Ishikawa T, Powers S, Kataoka T, Broek D, et al. In yeast, RAS proteins are controlling elements of adenylate cyclase. Cell. 1985 Jan 1; 40(1):27–36. https://doi.org/10.1016/0092-8674(85)90305-8 PMID: 2981630

85. Rolland F, De Winde JH, Lemaire K, Boles E, Thevelein JM, Winderickx J. Glucose-induced cAMP signalling in yeast requires both a G-protein coupled receptor system for extracellular glucose detection and a separable hexose kinase-dependent sensing process. Mol Microbiol. 2000 Oct; 38(2):348–58. https://doi.org/10.1046/j.1365-2958.2000.02125.x PMID: 11069660

86. Peeters K, Van Leemputte F, Fischer B, Bonini BM, Quezada H, Tsytlonok M, et al. Fructose-1,6-bisphosphate couples glycolytic flux to activation of Ras. Nat Commun. 2017 Dec 1; 8(1). https://doi.org/10.1038/s41467-017-01019-z PMID: 29030545

87. Toda T, Cameron S, Sass P, Zoller M, Wigler M. Three different genes in S. cerevisiae encode the catalytic subunits of the cAMP-dependent protein kinase. Cell. 1987 Jul 17; 50(2):277–87. https://doi.org/10.1016/0092-8674(87)90223-6 PMID: 3036373

88. Toda T, Cameron S, Sass P, Zoller M, Scott JD, McMullen B, et al. Cloning and characterization of BCY1, a locus encoding a regulatory subunit of the cyclic AMP-dependent protein kinase in Saccharomyces cerevisiae. Mol Cell Biol. 1987 Apr 1; 7(4):1371–7. https://doi.org/10.1128/mcb.7.4.1371 PMID: 3037314

89. Matsumoto K, Uno I, Toh-E A, Ishikawa T, Oshima Y. Cyclic AMP may not be involved in catabolite repression in Saccharomyes cerevisiae: evidence from mutants capable of utilizing it as an adenine source. J Bacteriol. 1982 Apr 1; 150(1):277–85. https://doi.org/10.1128/JB.150.1.277-285.1982 PMID: 6277865

90. Peeters T, Louwet W, Geladé R, Nauwelaers D, Thevelein JM, Versele M. Kelch-repeat proteins interacting with the Gα protein Gpa2 bypass adenylate cyclase for direct regulation of protein kinase A in yeast. Proc Natl Acad Sci U S A. 2006 Aug 29; 103(35):13034–9. https://doi.org/10.1073/pnas.0509644103 PMID: 16924114

91. Sass P, Field J, Nikawa J, Toda T, Wigler M. Cloning and characterization of the high-affinity cAMP phosphodiesterase of Saccharomyces cerevisiae. Proc Natl Acad Sci U S A. 1986; 83(24):9303–7. https://doi.org/10.1073/pnas.83.24.9303 PMID: 3025832

92. Nikawa J, Sass P, Wigler M. Cloning and characterization of the low-affinity cyclic AMP phosphodiesterase gene of Saccharomyces cerevisiae. Mol Cell Biol. 1987 Oct; 7(10):3629–36. https://doi.org/10.1128/mcb.7.10.3629 PMID: 2824992

93. Ma P, Wera S, Van Dijck P, Thevelein JM. The PDE1-encoded low-affinity phosphodiesterase in the yeast Saccharomyces cerevisiae has a specific function in controlling agonist- induced cAMP signaling. Mol Biol Cell. 1999; 10(1):91–104. https://doi.org/10.1091/mbc.10.1.91 PMID: 9880329

94. Hu Y, Liu E, Bai X, Zhang A. The localization and concentration of the PDE2-encoded high-affinity cAMP phosphodiesterase is regulated by cAMP-dependent protein kinase A in the yeast Saccharomyces cerevisiae. FEMS Yeast Res. 2010 Mar; 10(2):177–87. https://doi.org/10.1111/j.1567-1364.2009.00598.x PMID: 20059552

95. Swinnen E, Wanke V, Roosen J, Smets B, Dubouloz F, Pedruzzi I, et al. Rim15 and the crossroads of nutrient signalling pathways in Saccharomyces cerevisiae. Vol. 1, Cell Division. BioMed Central; 2006. p. 3. https://doi.org/10.1186/1747-1028-1-3 PMID: 16759348

96. Pedruzzi I, Bürckert N, Egger P, De Virgilio C. Saccharomyces cerevisiae Ras/cAMP pathway controls post-diauxic shift element-dependent transcription through the zinc finger protein Gis1. EMBO J. 2000 Jun 1; 19(11):2569–79. https://doi.org/10.1093/emboj/19.11.2569 PMID: 10835355

97. Martínez-Pastor MT, Marchler G, Schüller C, Marchler-Bauer A, Ruis H, Estruch F. The Saccharomyces cerevisiae zinc finger proteins Msn2p and Msn4p are required for transcriptional induction through the stress response element (STRE). EMBO J. 1996 May; 15(9):2227–35. PMID: 8641288

98. Schepers W, Van Zeebroeck G, Pinkse M, Verhaert P, Thevelein JM. In vivo phosphorylation of Ser21 and Ser83 during nutrient-induced activation of the yeast protein kinase A (PKA) target trehalase. J Biol Chem. 2012 Dec 28; 287(53):44130–42. https://doi.org/10.1074/jbc.M112.421503 PMID: 23155055

99. Dihazi H, Kessler R, Eschrich K. Glucose-induced stimulation of the Ras-cAMP pathway in yeast leads to multiple phosphorylations and activation of 6-phosphofructo-2-kinase. Biochemistry. 2003 May 27; 42(20):6275–82. https://doi.org/10.1021/bi034167r PMID: 12755632

100. Portela P, Howell S, Moreno S, Rossi S. In vivo and in vitro phosphorylation of two isoforms of yeast pyruvate kinase by protein kinase A. J Biol Chem. 2002 Aug 23; 277(34):30477–87. https://doi.org/10.1074/jbc.M201094200 PMID: 12063246

101. Rittenhouse J, Moberly L, Marcus F. Phosphorylation in vivo of yeast (Saccharomyces cerevisiae) fructose-1,6-bisphosphatase at the cyclic AMP-dependent site. J Biol Chem. 1987 Jul 25; 262 (21):10114–9. PMID: 3038868

102. Reinke A, Anderson S, McCaffery JM, Yates J, Aronova S, Chu S, et al. TOR Complex 1 Includes a Novel Component, Tco89p (YPL180w), and Cooperates with Ssd1p to Maintain Cellular Integrity in Saccharomyces cerevisiae. J Biol Chem. 2004 Apr 9; 279(15):14752–62. https://doi.org/10.1074/jbc.M313062200 PMID: 14736892

103. Dubouloz F, Deloche O, Wanke V, Cameroni E, De Virgilio C. The TOR and EGO protein complexes orchestrate microautophagy in yeast. Mol Cell. 2005 Jul 1; 19(1):15–26. https://doi.org/10.1016/j.molcel.2005.05.020 PMID: 15989961

104. Binda M, Péli-Gulli MP, Bonfils G, Panchaud N, Urban J, Sturgill TW, et al. The Vam6 GEF Controls TORC1 by Activating the EGO Complex. Mol Cell. 2009 Sep 11; 35(5):563–73. https://doi.org/10.1016/j.molcel.2009.06.033 PMID: 19748353

105. Bonfils G, Jaquenoud M, Bontron S, Ostrowicz C, Ungermann C, De Virgilio C. Leucyl-tRNA Synthetase Controls TORC1 via the EGO Complex. Mol Cell. 2012 Apr 13; 46(1):105–10. https://doi.org/10.1016/j.molcel.2012.02.009 PMID: 22424774

106. Bar-Peled L, Chantranupong L, Cherniack AD, Chen WW, Ottina KA, Grabiner BC, et al. A tumor suppressor complex with GAP activity for the Rag GTPases that signal amino acid sufficiency to mTORC1. Science (80-). 2013 May 31; 340(6136):1100–6.

107. Urban J, Soulard A, Huber A, Lippman S, Mukhopadhyay D, Deloche O, et al. Sch9 Is a Major Target of TORC1 in Saccharomyces cerevisiae. Mol Cell. 2007 Jun 8; 26(5):663–74. https://doi.org/10.1016/j.molcel.2007.04.020 PMID: 17560372

108. Yan G, Shen X, Jiang Y. Rapamycin activates Tap42-associated phosphatases by abrogating their association with Tor complex 1. EMBO J. 2006 Aug 9; 25(15):3546–55. https://doi.org/10.1038/sj.emboj.7601239 PMID: 16874307

109. Wanke V, Cameroni E, Uotila A, Piccolis M, Urban J, Loewith R, et al. Caffeine extends yeast lifespan by targeting TORC1. Mol Microbiol. 2008 Jul 1; 69(1):277–85. https://doi.org/10.1111/j.1365-2958.2008.06292.x PMID: 18513215

110. Beck T, Hall MN. The TOR signalling pathway controls nuclear localization of nutrient- regulated transcription factors. Nature. 1999 Dec 9; 402(6762):689–92. https://doi.org/10.1038/45287 PMID: 10604478

111. Kuruvilla FG, Shamji AF, Schreiber SL. Carbon- and nitrogen-quality signaling to translation are mediated by distinct GATA-type transcription factors. Proc Natl Acad Sci U S A. 2001 Jun 19; 98(13):7283–8. https://doi.org/10.1073/pnas.121186898 PMID: 11416207

112. Georis I, Feller A, Vierendeels F, Dubois E. The Yeast GATA Factor Gat1 Occupies a Central Position in Nitrogen Catabolite Repression-Sensitive Gene Activation. Mol Cell Biol. 2009 Jul 1; 29(13):3803–15. https://doi.org/10.1128/MCB.00399-09 PMID: 19380492

113. Liu Z, Butow RA. A Transcriptional Switch in the Expression of Yeast Tricarboxylic Acid Cycle Genes in Response to a Reduction or Loss of Respiratory Function. Mol Cell Biol. 1999 Oct 1; 19(10):6720–8. https://doi.org/10.1128/mcb.19.10.6720 PMID: 10490611

114. Dilova I, Aronova S, Chen JCY, Powers T. Tor signaling and nutrient-based signals converge on Mks1p phosphorylation to regulate expression of Rtg1p·Rtg3p-dependent target genes. J Biol Chem. 2004 Nov 5; 279(45):46527–35. https://doi.org/10.1074/jbc.M409012200 PMID: 15326168

115. Lempiäinen H, Uotila A, Urban J, Dohnal I, Ammerer G, Loewith R, et al. Sfp1 Interaction with TORC1 and Mrs6 Reveals Feedback Regulation on TOR Signaling. Mol Cell. 2009 Mar 27; 33(6):704–16.

116. Castermans D, Somers I, Kriel J, Louwet W, Wera S, Versele M, et al. Glucose-induced posttranslational activation of protein phosphatases PP2A and PP1 in yeast. Cell Res. 2012 Jun; 22(6):1058–77. https://doi.org/10.1038/cr.2012.20 PMID: 22290422

117. Nicastro R, Tripodi F, Gaggini M, Castoldi A, Reghellin V, Nonnis S, et al. Snf1 phosphorylates adenylate cyclase and negatively regulates protein kinase A-dependent transcription in Saccharomyces cerevisiae. J Biol Chem. 2015 Oct 9; 290(41):24715–26. https://doi.org/10.1074/jbc.M115.658005 PMID: 26309257

118. Orth JD, Thiele I, Palsson BØ. What is flux balance analysis? Nat Biotechnol. 2010; 28(3):245–8. https://doi.org/10.1038/nbt.1614 PMID: 20212490

119. Varma A, Palsson BO. Metabolic flux balancing: Basic concepts, scientific and practical use. Bio/Technology. 1994;

120. Famili I, Forster J, Nielsen J, Palsson BØ. Saccharomyces cerevisiae phenotypes can be predicted by using constraint-based analysis of a genome-scale reconstructed metabolic network. Proc Natl Acad Sci U S A. 2003 Nov; 100(23):13134–9. https://doi.org/10.1073/pnas.2235812100 PMID: 14578455

121. Wang M, Herrmann CJ, Simonovic M, Szklarczyk D, von Mering C. Version 4.0 of PaxDb: Protein abundance data, integrated across model organisms, tissues, and cell-lines. Proteomics. 2015 Sep 1; 15(18):3163–8. https://doi.org/10.1002/pmic.201400441 PMID: 25656970

122. Bateman A. UniProt: A worldwide hub of protein knowledge. Nucleic Acids Res. 2019;

123. Lewis NE, Hixson KK, Conrad TM, Lerman JA, Charusanti P, Polpitiya AD, et al. Omic data from evolved E. coli are consistent with computed optimal growth from genome-scale models. Mol Syst Biol. 2010;6. https://doi.org/10.1038/msb.2010.47 PMID: 20664636

124. Kauffman KJ, Prakash P, Edwards JS. Advances in flux balance analysis. Current Opinion in Biotechnology. 2003. https://doi.org/10.1016/j.copbio.2003.08.001 PMID: 14580578

125. Schuetz R, Kuepfer L, Sauer U. Systematic evaluation of objective functions for predicting intracellular fluxes in Escherichia coli. Mol Syst Biol. 2007; https://doi.org/10.1038/msb4100162 PMID: 17625511

126. Oliveira AP, Nielsen J, Förster J. Modeling Lactococcus lactis using a genome-scale flux model. BMC Microbiol. 2005; https://doi.org/10.1186/1471-2180-5-39 PMID: 15982422

127. Wang M, Herrmann CJ, Simonovic M, Szklarczyk D, von Mering C. Version 4.0 of PaxDb: Protein abundance data, integrated across model organisms, tissues, and cell-lines. Proteomics. 2015;

128. Wiśniewski JR, Rakus D. Multi-enzyme digestion FASP and the 'Total Protein Approach'-based absolute quantification of the Escherichia coli proteome. J Proteomics. 2014; https://doi.org/10.1016/j.jprot.2014.07.012 PMID: 25063446

129. Balakrishnan R, Park J, Karra K, Hitz BC, Binkley G, Hong EL, et al. YeastMine—an integrated data warehouse for Saccharomyces cerevisiae data as a multipurpose tool-kit. Database (Oxford). 2012; 2012:bar062. https://doi.org/10.1093/database/bar062 PMID: 22434830