MAURÍCIO ALEXANDER DE MOURA FERREIRA

MODEL-DRIVEN EVALUATION OF MICROBIAL PHYSIOLOGY: INSIGHTS FROM PROTEIN ALLOCATION

Thesis submitted to the Agricultural Microbiology Graduate Program of the Universidade Federal de Viçosa in partial fulfillment of the requirements for the degree of *Doctor Scientiae*.

Adviser: Wendel Batista da Silveira

Co-adviser: Zoran Nikoloski

VIÇOSA - MINAS GERAIS
2024

Ficha catalográfica elaborada pela Biblioteca Central da Universidade Federal de Viçosa - Campus Viçosa

T
F383m
2024

Ferreira, Maurício Alexander de Moura, 1995-
Model-driven evaluation of microbial physiology: insights from protein allocation / Maurício Alexander de Moura Ferreira. – Viçosa, MG, 2024.
1 tese eletrônica (221 f.): il. (algumas color.).

Inclui apêndices.
Orientador: Wendel Batista da Silveira.
Tese (doutorado) - Universidade Federal de Viçosa, Departamento de Microbiologia, 2024.
Inclui bibliografia.
DOI: https://doi.org/10.47328/ufvbbt.2024.425
Modo de acesso: World Wide Web.

1. Micro-organismos - Fisiologia. 2. Sistemas biológicos. 3. Engenharia metabólica. 4. Aprendizado do computador. I. Silveira, Wendel Batista da, 1979-. II. Universidade Federal de Viçosa. Departamento de Microbiologia. Programa de Pós-Graduação em Microbiologia Agrícola. III. Título.

CDD 22. ed. 579.562

Bibliotecário(a) responsável: Alice Regina Pinto Pires CRB-6/2523

MAURÍCIO ALEXANDER DE MOURA FERREIRA

MODEL-DRIVEN EVALUATION OF MICROBIAL PHYSIOLOGY: INSIGHTS FROM PROTEIN ALLOCATION

Thesis submitted to the Agricultural Microbiology Graduate Program of the Universidade Federal de Viçosa in partial fulfillment of the requirements for the degree of *Doctor Scientiae*.

APPROVED: 26 July 2024

Assent:

Documento assinado digitalmente
gov.br
MAURICIO ALEXANDER DE MOURA FERREIRA
Data: 30/07/2024 05:28:03-0300
Verifique em https://validar.iti.gov.br

Maurício Alexander de Moura Ferreira
Author

Documento assinado digitalmente
gov.br
WENDEL BATISTA DA SILVEIRA
Data: 30/07/2024 08:08:38-0300
Verifique em https://validar.iti.gov.br

Wendel Batista da Silveira
Adviser

ACKNOWLEDGEMENTS

My heartfelt gratitude:

To the Federal University of Viçosa, for the opportunity to complete the postgraduate course.

To the Agricultural Microbiology Graduate Program, for the opportunity to complete the postgraduate course.

To the University of Potsdam, for hosting me during the visiting PhD scholarship.

To the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES), for granting the scholarship in Brazil and for the DAAD/CAPES visiting PhD scholarship. This study was financed in part by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior – Brasil (CAPES) – Finance Code 001.

To Prof. Dr. Wendel Batista da Silveira, for the supervision.

To Prof. Dr. Zoran Nikoloski, for hosting me during the visiting PhD scholarship and for the co-supervision.

To Dr. Eduardo Luís Menezes de Almeida, Marius Arend and Philipp Wendering, for the contributions to this thesis.

To Prof. Dr. Leonardo Lopes Bhering, for granting access to his lab's Linux server.

To all my colleagues at the Fisiologia de Microrganismos laboratory (Department of Microbiology, Federal University of Viçosa), for all contributions, discussions and friendship.

To all my colleagues at the Bioinformatics Group (Institute of Biochemistry and Biology, University of Potsdam), for all contributions, discussions and friendship.

“There is no royal road to science, and only those who do not dread the fatiguing climb of its steep paths have a chance of gaining its luminous summits.”

(Karl Marx, Capital: A Critique of Political Economy, Volume 1)

ABSTRACT

FERREIRA, Maurício Alexander de Moura, D.Sc., Universidade Federal de Viçosa, July, 2024. **Model-driven evaluation of microbial physiology: insights from protein allocation**. Adviser: Wendel Batista da Silveira. Co-adviser: Zoran Nikoloski.

The optimal allocation of proteins to cellular functions is crucial for cell survival and growth. However, the strategies employed by the cell are still elusive, as there are many supposedly conflicting objectives to be considered, such as minimizing the expenditure of resources, while at the same time affording to produce certain enzymes in excess, despite the lower demand for enzyme resources to maintain a certain amount of metabolic flux. Further, certain phenotypes, such as the overflow metabolism, are triggered by changes in resource distribution. In order to tackle these problems, the thesis focuses on the usage of protein-constrained metabolic models in combination with machine learning and integration with multi-omics data. Based on these approaches, here it is predicted the occurrence of overflow metabolism in the form of respiro-fermentative metabolism in the yeast *Kluyveromyces marxianus*. By integrating the metabolic model of *K. marxianus* with transcriptomics data, new insights on the genes, enzymes and metabolites involved in ethanol stress were obtained. Next, it is presented a new approach for studying enzyme usage redistribution, PARROT, which minimizes the distance between enzyme usage of an initial growth condition and a changing growth condition, based on the principle of minimal adjustment. The PARROT approach was able to predict enzyme usage in alternative growth conditions with higher accuracy than previous methods. While this approach is useful for studying resource redistribution, it is still not able to predict *in vivo* protein concentrations, given that the predicted usage is limited to a given metabolic flux and catalytic efficiency. To solve this problem, an approach that combines machine learning with metabolic modelling was developed, termed CAMEL. This approach could accurately predict *in vivo* concentrations, including for strains that were metabolically engineered. Finally, resource redistribution was evaluated on the context of enzyme promiscuity, from which a network of reactions termed “underground metabolism” can arise. To this end, the approach named CORAL was developed to

integrate enzyme promiscuity constraints into metabolic models. It was found that these promiscuous enzymes are important for maintaining growth and providing robustness to disturbances in metabolism. The results obtained in this thesis are relevant to systems metabolic engineering endeavours, providing tools and knowledge to design microbial strains more suitable for industrial applications.

Keywords: Systems biology; Metabolic engineering; Microbial physiology; Machine learning.

RESUMO

FERREIRA, Maurício Alexander de Moura, D.Sc., Universidade Federal de Viçosa, julho de 2024. **Model-driven evaluation of microbial physiology: insights from protein allocation**. Orientador: Wendel Batista da Silveira. Coorientador: Zoran Nikoloski.

A alocação ótima de proteínas para as funções celulares é essencial para a sobrevivência e crescimento da célula. No entanto, as estratégias empregadas pela célula ainda são pouco compreendidas, já que há objetivos conflitantes a serem considerados, como minimizar o gasto de recursos e, ao mesmo tempo, permitir a produção de determinadas enzimas em excesso, apesar da menor demanda por recursos enzimáticos para manter uma determinada quantidade de fluxo metabólico. Além disso, determinados fenótipos, como o metabolismo *overflow*, são desencadeados por mudanças na distribuição de proteínas. Para resolver esses problemas, esta tese se concentra no uso de modelos metabólicos com parâmetros enzimáticos em combinação com aprendizado de máquina e integração com dados multiômicos. Com base nessas abordagens, foi possível predizer a ocorrência do metabolismo *overflow* na forma de metabolismo respiro-fermentativo na levedura *Kluyveromyces marxianus*. Ao integrar o modelo metabólico da *K. marxianus* com dados de transcriptômica, foram obtidas novas informações sobre os genes, enzimas e metabólitos envolvidos no estresse causado por etanol. Em seguida, é apresentada uma nova abordagem para o estudo da redistribuição do uso de enzimas, PARROT, que minimiza a distância entre uma condição de crescimento inicial e uma condição de crescimento alternativa, com base no princípio do ajuste mínimo. A abordagem PARROT foi capaz de predizer o uso de enzimas em condições alternativas de crescimento com maior acurácia do que outros métodos. Embora essa abordagem seja útil para estudar a redistribuição de recursos, ela não é capaz de predizer as concentrações de proteínas *in vivo*, uma vez que o uso previsto é limitado a um determinado fluxo metabólico e uma dada eficiência catalítica. Para resolver esse problema, foi desenvolvida uma abordagem que combina aprendizado de máquina com modelagem metabólica, denominada CAMEL, que pode predizer com acurácia

as concentrações *in vivo* de enzimas, inclusive para linhagens submetidas à engenharia metabólica. Por fim, a redistribuição de recursos foi avaliada no contexto da promiscuidade enzimática, a partir da qual pode se formar uma rede de reações denominada "metabolismo *underground*". Para tal, foi desenvolvida a abordagem denominada CORAL para integrar os parâmetros de promiscuidade enzimática a modelos metabólicos. Evidenciou-se que as enzimas promíscuas são importantes para manter o crescimento e proporcionar robustez a perturbações no metabolismo. Os resultados obtidos nesta tese também são relevantes para o avanço de estratégias de engenharia metabólica sistêmica, fornecendo ferramentas e conhecimento para construir linhagens microbianas mais apropriadas para aplicações industriais.

Palavras-chave: Biologia de sistemas; Engenharia metabólica; Fisiologia microbiana; Aprendizado de máquina.

SUMMARY

1 INTRODUCTION

Proteins are the workhorses of cellular form and function, being responsible for catalysing reactions in metabolism, transporting molecules across the cytosol, importing and exporting molecules from the cell, signalling environmental changes, and providing structure to the cell (NELSON; COX, 2021).

Each individual protein is present in multiple copies in the cell. The number of copies of a protein (i.e. protein abundance) relies on a series of factors, primarily by balancing protein biosynthesis and protein degradation, but also by the regulation of transcription and translation, codon usage bias, and protein modifications (LIU; BEYER; AEBERSOLD, 2016). Each copy of a certain protein takes cell resources to be synthesized, such as amino acids and ATP. For example, the incorporation of one amino acid requires around 4.5 molecules of ATP by the ribosome machinery (ZERIHUN; MCKENZIE; MORTON, 1998). Besides, each molecular function inside the cell requires a certain number of copies of a certain protein in order to work properly. The protein abundance is condition-dependent in cells; thus, the fine-tuning of protein abundance is important for proper cellular function and survival (LIU; BEYER; AEBERSOLD, 2016). As such, cells then face an optimization problem: how many protein copies should it allocate to its functions, while maximizing growth and minimizing resource usage?

The collection of all protein molecules in the cell in a given circumstance – the cellular proteome – must be carefully allocated to each cellular function according to its necessity in order to guarantee appropriate cell growth and metabolism (BERTAUX; MARGUERAT; SHAHREZAEI, 2018). However, despite knowing its mechanisms, it is still not understood how the exact protein stoichiometry is determined. It was observed that the preferred range of enzyme stoichiometry follows a narrow distribution among pathways in Gram-positive and -negative bacteria, likely a result of evolutionary conservation or convergence. The cellular proteome is highly condition-dependent, and therefore moulded by environmental factors and growth conditions. When the conditions are unstable, the cell must be able to quickly adapt to these unstable conditions, or else perish. These changing conditions trigger a series of signalling pathways that culminate in the alteration of gene expression, leading to changes in protein abundance (HUI et al., 2015). Nevertheless, it seems that many proteins are

paradoxically produced in a quantity larger than necessary, especially enzymes in metabolic reactions. This over-abundance of enzymes could act as a buffer to allow the cell to quickly adapt to variation in nutrient uptakes or other changes in the environment (MORI et al., 2017). Given the apparently conflicting objectives for the problem of resource allocation in changing environments, what strategy is employed by the cell to ensure that the cellular proteome is correctly adjusted? Further, what is, then, the implications of this disparity between the abundance of an enzyme and its actual usage for the problem of resource allocation?

One interesting phenotype shift associated with protein allocation adjustments is the overflow metabolism – where the cell redirects metabolic flux from respiration to fermentation, despite having sufficient oxygen available. In this scenario, the high concentration of the carbon source causes the enzymes from the respiratory pathways to be stalled, while enzymes from fermentative pathways are induced. This strategy is based on the lower Gibbs energy dissipation of fermentative pathways (NIEBEL; LEUPOLD; HEINEMANN, 2019), as well as the lower resource cost of enzymes in these pathways (BASAN et al., 2015). However, the products of fermentative pathways can also be toxic to the cell, such as ethanol (DE MOURA FERREIRA; DA SILVEIRA; DA SILVEIRA, 2022). While the mechanisms for dealing with these conditions are well understood in model species such as *Escherichia coli* and *Saccharomyces cerevisiae*, it raises the question of how non-model species behave, which also display potential for applications in bioprocesses.

This thesis sheds light on the many questions brought up in this introduction, dealing with the problem of resource allocation in microbial and cell physiology, specifically enzyme resources across metabolism. A better understanding of the issues raised herein are not important only from a microbial physiology point of view, but also in terms of systems metabolic engineering approaches. It is important to point out that these approaches are considered pivotal to improve the construction of robust microbial strains with enhanced capacity to produce biobased-products in biorefineries, which is mandatory to reduce the petrochemical industry-dependence. The thesis is divided into two sections. In the first section, the physiology of the yeast *Kluyveromyces marxianus* is investigated focusing on enzyme resource distribution in shifting phenotypes. Chapter 2.1 presents a robust review of *K. marxianus*, its responses to ethanol stress and the standing questions regarding its physiology.

Chapter 2.2 assesses resource distribution and the adaptations to ethanol stress. In the second section, the modelling capabilities of pcGEMs are leveraged to assess enzyme usage more thoroughly, focusing on the model organisms *Escherichia coli* and *Saccharomyces cerevisiae*. Chapter 3.1 introduces and reviews protein-constrained genome-scale metabolic models (pcGEMs). In Chapter 3.2, it is addressed the open question of how cells reorganize the proteome in response to different environmental conditions. Later, Chapter 3.3 investigates how the cells organize a "protein reserve" to tackle changing conditions. Finally, Chapter 3.4 concerns with how enzyme resources are used and distributed across promiscuous enzymes, including the so-called "underground" metabolic activity that can arise from such promiscuous enzymes.

References

BASAN, M. et al. Overflow metabolism in Escherichia coli results from efficient proteome allocation. **Nature**, v. 528, n. 7580, p. 99–104, 2015.

BERTAUX, F.; MARGUERAT, S.; SHAHREZAEI, V. Division rate, cell size and proteome allocation: Impact on gene expression noise and implications for the dynamics of genetic circuits. **Royal Society Open Science**, v. 5, n. 3, 2018.

DE MOURA FERREIRA, M. A.; DA SILVEIRA, F. A.; DA SILVEIRA, W. B. Ethanol stress responses in Kluyveromyces marxianus: current knowledge and perspectives. **Applied Microbiology and Biotechnology**, v. 106, n. 4, p. 1341–1353, 1 fev. 2022.

HUI, S. et al. Quantitative proteomic analysis reveals a simple strategy of global resource allocation in bacteria. **Molecular Systems Biology**, v. 11, n. 2, p. 784, 2015.

LIU, Y.; BEYER, A.; AEBERSOLD, R. On the Dependency of Cellular Protein Levels on mRNA Abundance. **Cell**, v. 165, n. 3, p. 535–550, 21 abr. 2016.

MORI, M. et al. Quantifying the benefit of a proteome reserve in fluctuating environments. **Nature Communications 2017 8:1**, v. 8, n. 1, p. 1–8, 31 out. 2017.

NELSON, D. L.; COX, M. **Lehninger Principles of Biochemistry: International Edition**. [s.l.] Macmillan Learning, 2021.

NIEBEL, B.; LEUPOLD, S.; HEINEMANN, M. An upper limit on Gibbs energy dissipation governs cellular metabolism. **Nature Metabolism**, v. 1, n. 1, p. 125–132, 2019.

ZERIHUN, A.; MCKENZIE, B. A.; MORTON, J. D. Photosynthate costs associated with the utilization of different nitrogen–forms: influence on the carbon balance of plants and shoot–root biomass partitioning. **The New Phytologist**, v. 138, n. 1, p. 1–11, jan. 1998.

2 METABOLIC SHIFTS AND STRESS RESPONSES IN *KLUYVEROMYCES MARXIANUS*

2.1 Ethanol stress responses in *Kluyveromyces marxianus*: current knowledge and perspectives

Text adapted from the unmarked revised manuscript as accepted by the Applied Microbiology and Biotechnology journal, available at https://doi.org/10.1007/s00253-022-11799-0.

Abstract

The rising concern with the emission of greenhouse gases has boosted new incentives for biofuels production, which are less polluting than fossil fuels. Special attention has been given to the second generation ethanol, as it is produced from abundant feedstocks which do not compete with food production, such as lignocellulosic biomass and whey. *Kluyveromyces marxianus* stands out in second generation ethanol production due to its capacity of assimilating lactose, the sugar found in whey, and tolerating high temperatures used in simultaneous saccharification processes. Nonetheless, contrary to *Saccharomyces cerevisiae*, *K. marxianus* does not tolerate high ethanol concentrations. Ethanol causes a broad range of toxic effects on yeasts, acting on cell membrane and proteins, as well as inducing the generation of reactive oxygen species (ROS). The ethanol stress responses are not fully understood, mainly in non-conventional yeasts such as *K. marxianus*. Indeed, many molecular responses to ethanol stress are still inferred from *S. cerevisiae*. As such, a better understanding of the ethanol stress responses in *K. marxianus* may provide the basis for improving its use in the biofuel industry. Additionally, the selection of ethanol-tolerant strains by metabolic engineering is useful to provide strains with improved capacity to withstand stressful conditions, as well as to obtain new insights about the ethanol stress responses.

Keywords

Yeasts, ethanol tolerance, stressful conditions, metabolic engineering, second generation ethanol.

Key points

- It is still not totally clear why *K. marxianus* is less tolerant to ethanol than *S. cerevisiae*.
- Understanding the ethanol stress response in *K. marxianus* is pivotal for improving its application in the biofuel industry.
- Metabolic engineering is expected to improve the ethanol tolerance in *K. marxianus*.

Introduction

The growing demand for fuels is a matter of great concern, as the use of fossil fuels are related to the emission of greenhouse gases. Recent studies showed that if the energetic demand continues to increase, the global temperature will increase 1.65 °C (IEA et al. 2021). As such, it is expected that new incentives boost the production of biofuels in the next years, as they are less polluting, (IEA et al. 2020). Therefore, the improvement and the development of new technologies of biofuel production is pivotal to meet the energetic demand in a more environmentally sustainable manner. Ethanol is still the main biofuel produced in the world. According to the Renewable Fuels Association (RFA), the world ethanol production hit a new record of 29,100 million gallons in 2020. The United States and Brazil are the world leaders of ethanol fuel (IEA et al. 2020). First generation ethanol uses carbohydrates from sugarcane, cereal crops and seeds as feedstock for ethanol production. Nevertheless, these feedstocks can compete with animal feed, human food, arable lands, fresh water and fertilizer requirements (Limayem and Ricke 2012). To mitigate these drawbacks, second generation ethanol, which is produced from lignocellulosic biomass and whey, has been exploited (Balat and Balat 2009; Naik et al. 2010; Sims et al. 2010).

Saccharomyces cerevisiae is widely employed for first ethanol generation production due to its ability of efficiently converting glucose and sucrose obtained from corn and sugarcane, respectively, into ethanol and dioxide carbon (Piškur et al. 2006; Stanley et al. 2010; Dequin and Casaregola 2011). However, this yeast is not capable of assimilating carbon sources such as lactose and xylose, which are found in whey and lignocellulosic biomass, respectively. In addition, *S. cerevisiae* does not tolerate high temperatures, which are required for ethanol production from lignocellulosic biomass using the simultaneous saccharification and fermentation (SSF) process.

Otherwise, *Kluyveromyces marxianus* strains are thermotolerant and capable of using lactose and xylose as the sole carbon sources (Kurtzman and Fell 1998; Fonseca et al. 2008; Fonseca et al. 2013). As such, *K. marxianus* displays some advantages over *S. cerevisiae* for second generation ethanol production. Over the last decades, several works underscored the potential of different *K. marxianus* strains for ethanol production from various feedstocks, including lignocellulosic biomass and whey (Table 1). Nonetheless, *K. marxianus*, in contrast to *S. cerevisiae* strains, does not tolerate high ethanol concentrations (Silveira et al. 2005; Diniz et al. 2017; Alvim et al. 2019; Silveira et al. 2020). This undesirable feature limits its use in bioprocesses with cell recycling, as cell viability is compromised at high ethanol concentrations.

Table 1. Ethanol production by several *K. marxianus strains* from different feedstocks.

Strain	Feedstock/substrate	Reference
K. marxianus CCT 7735 (UFV-3)	cheese whey permeate	(Silveira et al. 2005) (Diniz et al. 2013)
	sugarcane bagasse and ricotta whey	(Ferreira et al. 2015)
	sugarcane bagasse	(de Souza et al. 2012)
	sugarcane bagasse and glucose	(Costa et al. 2014)
	elephant grass	(Campos et al. 2019)
	Yeast-Nitrogen-Base medium with glucose/xylose solution	(Dos Santos et al. 2013)
	sweet sorghum bagasse	(Tinôco et al. 2021)
K. marxianus (own isolates)	overripe mango pulp	(Buenrostro-Figueroa et al. 2018)
K. marxianus (own isolates)	cheese whey	(Hesham et al. 2014)
K. marxianus BY25569	rice waste biomass	(Saratale et al. 2017)

K. marxianus CBS1555 (KCTC7001)	empty palm fruit bunches	(Jung et al. 2015)
K. marxianus DBKKUY-103	sorghum juice	(Pilap et al. 2018)
K. marxianus DMB3-7	Yeast-Nitrogen-Base medium with xylose solution	(Goshima et al. 2013)
K. marxianus IFO 0288	cheese whey	(Koutinas et al. 2014)
K. marxianus K21	taro waste	(Wu et al. 2016)
K. marxianus KD-15	saccharified flour mixed with cheese whey saccharified potato tubers mixed with cheese whey	(Nakamura et al. 2012)
K. marxianus LOCK0024	tomato, pepper, grape, and acid whey	(Güneşer et al. 2016)
K. marxianus TISTR5925 (tryptophan auxotroph)	cassava pulp hydrolysates	(Rugthaworn et al. 2014)
K. marxianus UMIP2234.94	Yeast-Malt-Dextrose with glucose solution	(Mehmood et al. 2018)
K. marxianus var. *marxianus* CBS 397	ricotta cheese whey, raw cheese whey and deproteinized whey	(Sansonetti et al. 2009)
K. marxianus var. *marxianus* CBS 712	cheese whey and ricotta cheese whey effluents	(Zoppellari and Bardi 2013)
K. marxianus Y01070	old corrugated cardboard and paper sludge	(Kádár et al. 2004)
K. marxianus Y179	Jerusalem artichoke tuber meal	(Yuan et al. 2012)
K. marxianus CECT 10875	wheat straw	(Moreno et al. 2013)
K. marxianus CECT 10875	wheat straw	(Tomás-Pejó et al. 2009)
K. marxianus K213	water hyacinth	(Yan et al. 2015)
	pretreated carrot pomace	(Yu et al. 2013)
K. marxianus OFF1	wheat straw	(Sandoval-Nuñez et al. 2018)

K. marxianus SLP1	sugarcane bagasse	(Sandoval-Nuñez et al. 2018)
K. marxianus DMKU 3-1042	sugarcane juice	(Limtong et al. 2007)
K. marxianus URM 7404	cheese whey	(Tavares et al. 2019)
Kluyveromyces sp. IIPE453	pretreated sugarcane bagasse pith	(Dasgupta et al. 2013)

Most of the studies involving ethanol stress responses in yeasts focused on *S. cerevisiae*; however, such responses are still not fully understood. So far, few studies with *K. marxianus* under ethanol stress have been conducted. Over the last few years, some studies analyzed the ethanol stress responses in *K. marxianus*, pointing out some differences between it and *S. cerevisiae*. These differences provide cues to why *K. marxianus* is less tolerant to ethanol. However, many responses known in *S. cerevisiae* have not yet been identified in *K. marxianus*, which presents a gap for future research to fill.

Therefore, this review focuses on the ethanol effect on *K. marxianus* physiology, mainly the damages caused by ethanol stress by outlining how ethanol interacts and disrupts cellular components and processes, and then detailing the molecular responses associated with each target of ethanol in the cell. Further, we explore recent metabolic engineering strategies to select *K. marxianus* ethanol-tolerant strains

Ethanol tolerance

Obtaining high ethanol titers is pivotal to reduce the distillation costs and make the process of bioethanol production economically feasible (Deparis et al. 2017). Nevertheless, high ethanol concentrations impose a stress condition that can impair the yeast growth or even lead to cell death. *S. cerevisiae* industrial strains such as PE 2 and CA1185 tolerate high ethanol concentrations, as more than 85% of cells are still viable up to 140 $g.L^{-1}$ of ethanol (Pereira et al. 2011). Otherwise, *K. marxianus* strains are less tolerant to high ethanol concentrations. Silveira et al. (2005) reported that the growth of *K. marxianus* CCT 7735 in cheese whey permeate medium was severely impaired at ethanol concentrations above 50 $g.L^{-1}$). Costa et al. (2014) demonstrated

that the strains of *K. marxianus* CCT 7735, ATCC 8554 and CCT 4086 were not tolerant to ethanol concentration of 6% (v/v) in YPD medium [2 % yeast extract, 1 % peptone, and 2 % glucose (w/v)]. Recently, an ethanol inhibition model for *K. marxianus* CCT 7735 was developed (Tinôco et al. 2021). In glucose concentrations superior to 20 $g.L^{-1}$, the Ghose & Tyagi-based model shows that the growth rate decreases in growing ethanol concentrations.

K. marxianus UCD (FST) 55-82 cultivated in a culture medium containing 10% (w/v) inulin-type sugars had its growth strongly inhibited in the presence of ethanol 8% (v/v) since the specific growth rate was reduced from 0.42 (control, absence of ethanol) to 0.09 h^{-1} (Bajpai and Margaritis 1982). Madeira-Jr and Gombert (2018) evaluated the ability of twenty *K. marxianus* strains to withstand high temperatures and ethanol concentrations. Overall, the growth of these strains was impaired in ethanol concentrations ranging from 5% to 10% (v/v) at 30 ºC in YPD medium.

Therefore, the low ethanol tolerance displayed by *K. marxianus*, in comparison to *S. cerevisiae*, is the main drawback for its application in the biofuel industry.

Ethanol stress

Ethanol has a chaotropic property, entropically disordering macromolecular structures (Cray et al. 2013; Eardley and Timson 2020). As such, ethanol causes a broad range of toxic effects on yeast cells, severely handicapping growth and cellular function (Navarro-Tapia et al. 2017). It interacts mainly with cell membranes, triggering the disruption of electrochemical gradients, membrane transport, metabolic and signaling pathways; and proteins, causing their denaturation and loss of function (Figure 1).

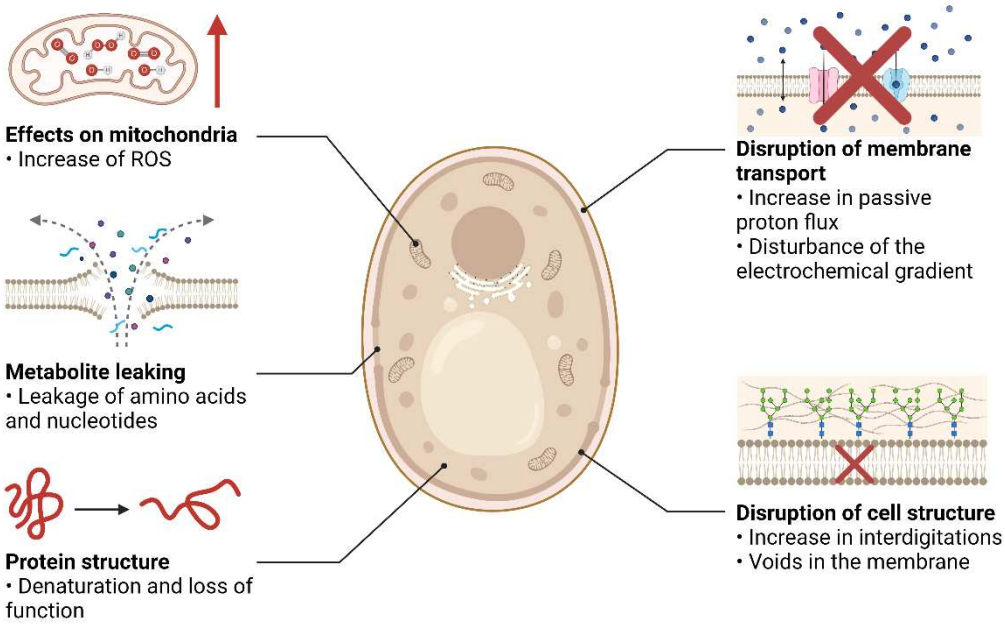

Fig. 1 Effects of ethanol on yeast. Created with BioRender.com

The effects of ethanol on cell membranes are due to the undermining of their integrity. Ethanol gets intercalated in the hydrophilic interior of the lipid bilayer by taking the place of water molecules. This increases lipid interdigitations, as the hydroxyl group of ethanol interacts with membrane head groups, increasing surface area and creating voids in the membrane. Then, these voids are filled with opposite monolayer groups (Devanand et al. 2019). The damage caused to membrane structure increases the passive proton flux through the membrane, leading to the disruption of the electrochemical gradient, which is essential to nutrient uptake. This causes a decrease in intracellular pH and depolarization of the plasma membrane, which in turn impairs nutrient transport (Charoenbhakdi et al. 2016; Li et al. 2018). The damage caused to the plasma membrane also leads to a leakage of metabolites to the extracellular medium, such as amino acids and nucleotides, decreasing growth (Silveira et al. 2020; Jhariya et al. 2021).

Proteins are another important target of ethanol. It affects protein structure and function by a series of mechanisms, such as disrupting hydrogen bonds between amino acid residues, competing with water for the solvation of the protein's surface, and interacting with hydrophobic residues, leading to the disruption of the hydrophobic core in globular proteins (Buck 1998; Martin et al. 2008; Nikolaidis et al. 2017).

The endogenous generation of reactive oxygen species (ROS), such as the highly reactive hydroxyl radical, hydrogen peroxide and superoxide, is another cellular process disturbed by ethanol. One source of endogenous ROS is the mitochondrial electron transport chain, which naturally produces ROS as byproducts of its function. However, ethanol impairs iron homeostasis in the mitochondria, the biogenesis and recycling of iron-sulfur clusters (Vamvakas and Kapolos 2020). Iron homeostasis is a tightly regulated process in mitochondria, and its disturbance leads to free iron and increase in ROS formation (Gomez et al. 2014). Ethanol also disturbs the mitochondrial membranes, reducing the proton motive forces and causing proton leakage (Yang et al. 2012).

Molecular responses to ethanol stress

Besides acting at specific targets, ethanol triggers many stress response mechanisms in yeasts in order to counteract the effects caused by its presence (Figure 2). Ethanol also induces oxidative stress responses in yeasts (Zhao and Bai 2009; Ma and Liu 2010; Stanley et al. 2010). These stress responses include the modulation of many cell processes, gene regulatory pathways, and changes in metabolite concentrations. *K. marxianus* has mechanisms to counteract the effects of ethanol stress; however, many molecular responses to ethanol stress are still not fully known, and much is still inferred from *S. cerevisiae* (Table 2) or its sister species, *Kluyveromyces lactis* (Ortiz-Merino et al. 2018).

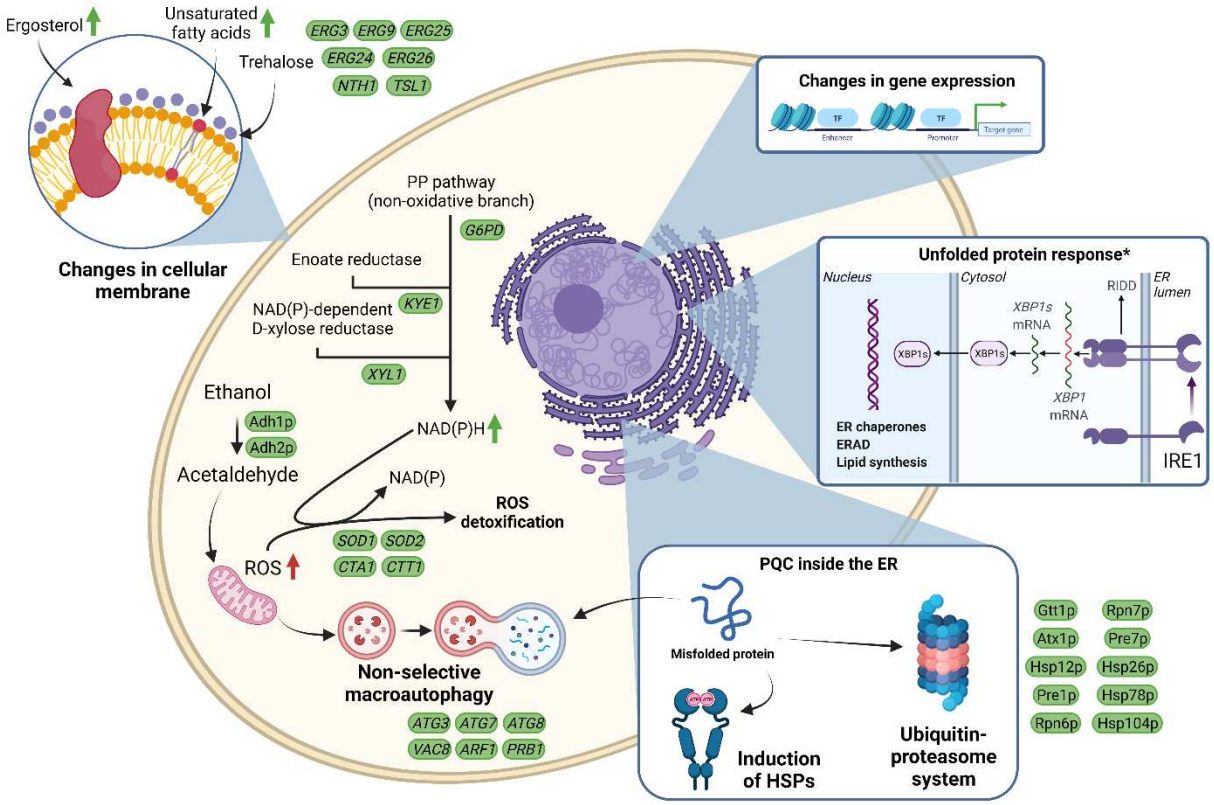

Fig. 2 Molecular responses to ethanol stress in *K. marxianus*. Green arrows indicate cellular responses. Red arrows indicate ethanol effects. Created with BioRender.com

Table 2 Comparison of known molecular responses between *S. cerevisiae* and *K. marxianus*.

Molecular responses	*Saccharomyces cerevisiae*	*Kluyveromyces marxianus*
Responses to cell membrane damage	• Increase of ergosterol and unsaturated fatty acids contents (Caspeta et al. 2014) • Accumulation of trehalose (Bandara et al. 2009)	• Alterations in ergosterol and unsaturated fatty acids contents is cultivation-dependent response (Fu et al. 2019) • Accumulation of trehalose (Alvim et al. 2019)
Responses to protein misfolding and denaturing	• Up-regulation of molecular chaperones (Saini et al. 2018)	• Up-regulation of molecular chaperones (Li et al. 2019, Alvim et al. 2019)

	• Sequestering of misfolded proteins in quality control compartments (Chen et al. 2011, Ho et al. 2019) • Activation of the UPS (Yoshida et al. 2021) • Activation of the UPR (Halbleib et al. 2017; Navarro-Tapia et al. 2017) • Induction of autophagy (Reggiori and Klionsky 2013)	• Activation of the UPS (Li et al. 2019, Alvim et al. 2019) • Induction of autophagy (Gao et al. 2015, Fu et al. 2019)
Oxidative stress caused by ethanol and its associated response	• Production of additional redox and energy cofactors (Yang et al. 2012) • Increased flux through the PPP (Celton et al. 2012) • Induction of autophagy (Reggiori and Klionsky 2013) • Detoxification of ROS (Yang et al. 2012)	• Production of additional redox and energy cofactors (Alvim et al. 2019, Schabort et al. 2016, Wang et al. 2018, Diniz et al. 2017) • Up-regulation of genes encoding enzymes related to NAD(P)H generation (Alvim et al. 2019, Schabort et al. 2016, Wang et al. 2018, Diniz et al. 2017) • Induction of autophagy (Gao et al. 2015, Fu et al. 2019) • Detoxification of ROS (Wang et al. 2018)

Metabolite concentration changes in response to ethanol stress	• Accumulation of glycerol (Vriesekoop et al. 2009) • Increased production of acetate (Stanley et al. 2010) • Increased concentration of alanine and methionine, decreased concentration of proline, isoleucine, threonine, aspartate, glutamate (Ming et al. 2019)	• Increased production of acetate (Fu et al. 2019) • Initial decrease in the concentration of arginine, proline, glutamate, phenylalanine and tryptophan, later increase at longer exposure to ethanol (Alvim et al. 2019)

Responses to cell membrane damage

Changes in lipid metabolism, especially membrane lipids, is an approach undertaken by the cell to mitigate damage caused by ethanol. This affects the metabolism of unsaturated fatty acids, the structural components of cell membranes; and of sterols, which affects the cell membrane fluidity and stability, and is one of the contributing factors to thermotolerance in yeasts (Caspeta et al. 2014). *S. cerevisiae* exposed to ethanol displays an increase of both ergosterol and unsaturated fatty acid contents, leading to the remodeling of its cell membrane. In *K. marxianus*, the changes in ergosterol metabolism in response to ethanol seems to be related to the stress condition evaluated. In the study of Fu et al. (2019), using *K. marxianus* DMKU3-1042 in stationary growth after fermenting ethanol, they detected an up-regulation of the genes *ERG3*, *ERG9*, *ERG25*, *ERG24*, and *ERG26*. This was correlated with an increase in ergosterol content as measured in the study. Contrarily, Diniz et al (2017) observed that both ergosterol and unsaturated fatty acids biosynthetic genes in *K. marxianus* CCT 7735 were down-regulated, and the intracellular content of unsaturated fatty acids and ergosterol did not change after ethanol stress. Therefore, the expression of genes encoding proteins involved with ergosterol biosynthesis increases when *K. marxianus* is exposed to rising ethanol concentrations over batch

cultivation, contrary to sudden exposure of exponentially growing cells to high ethanol concentration.

Kosmotropic agents are a useful class of molecules that can help the cell counteract the effects of ethanol. These molecules have the opposite effect of chaotropic agents; they help macromolecules be more stable in water solutions (Cray et al. 2013). The eukaryotic cell possesses many such molecules, which have their biosynthetic pathways induced by the presence of ethanol or of other stress signaling molecules. One such molecule is the disaccharide trehalose, which has properties such as being highly hydrophilic, highly stable and lacking internal hydrogen bonding (Saini et al. 2018). It acts as a chemical co-chaperone, binding to other chaperones such as heat-shock proteins (HSPs), and thus helps keep proteins from being misfolded. It can also help stabilize the cell membrane, by displacing ethanol from the membrane and taking its place instead (Ding et al. 2009; Šoštarić et al. 2021). In *S. cerevisiae*, accumulation of trehalose is known to be an important contributor to cell survival in high ethanol concentrations (Bandara et al. 2009). In *K. marxianus* CCT 7735, Alvim et al. (2019) found that after 4 h of ethanol exposure, trehalose levels sharply increased. A transcriptomics analysis performed by Mo et al. (2019) of *K. marxianus* growing in 6% ethanol demonstrated that the genes *NTH1* and *TSL1*, involved in trehalose biosynthesis and degradation, were up-regulated.

Responses to protein misfolding and denaturing

Under ethanol stress, the protein quality control (PQC) system is affected as a result of protein denaturation and loss of function. The PQC is pivotal in mitigating damage caused by misfolded proteins, and it employs a series of mechanisms such as up-regulating molecular chaperones, activating the ubiquitin-proteasome system (UPS), or inducing autophagy (Yoshida et al. 2021). In *S. cerevisiae*, the production of HSPs is highly induced, such as Hsp104p, Hsp82p, Hsp70p, Hsp26p, Hsp30p and Hsp12p (Saini et al. 2018). Further, Hsp42p and Btn2p act in sequestering misfolded proteins and preventing damage to the cell (Chen et al. 2011). During stress conditions, the protein Btn2p also helps protein folding by recruiting the Hsp70-Hsp104 disaggregase and redirecting sequestered proteins to the refolding pathway (Ho et al. 2019). A proteomic profiling integrated with transcriptomics conducted by Li et al. (2019), using the strain *K. marxianus* DMKU3-1042, detected an up-regulation of several molecular

chaperones, some of which are Gtt1p, Atx1p and Hsp12p. They also detected an up-regulation of many proteasome proteins, such as Pre1p, Rpn6p, Rpn7p and Pre7p. Similarly, Alvim et al (2019) detected an up-regulation of Hsp26p, Hsp78p and Hsp104p in the strain *K. marxianus* CCT 7735.

Another important mechanism to deal with unfolded proteins is the unfolded protein response (UPR), a signaling pathway that regulates protein homeostasis in the endoplasmic reticulum (ER) (Halbleib et al. 2017; Navarro-Tapia et al. 2017). The protein Ire1p, which is a transmembrane protein embedded in the ER membrane, is key to activating UPR in yeasts (Adams et al. 2019). The UPR is activated mainly in response to accumulation of unfolded proteins in the ER due to stressful conditions, and works by reducing protein translation, degrading misfolded proteins and inducing protein folding (Read and Schröder 2021). It is therefore of utmost importance to minimize the impact of unfolded or misfolded proteins in the cell, as failure to do so results in cell death. In *S. cerevisiae*, changes in membrane fluidity can also trigger UPR (Halbleib et al. 2017; Navarro-Tapia et al. 2017). Given that the ER is also a site for lipid metabolism, the UPR also takes part in lipid homeostasis, mainly through the X-box binding protein 1 (XBP1) and regulated inositol-requiring enzyme 1 (Ire1)-dependent decay (RIDD) pathways, assisting the cell in recovering and maintaining membrane structural integrity (Moncan et al. 2021). In *Schizosaccharomyces pombe*, approximately 31% of the genes regulated by Ire1p are related to lipid metabolism (Hernández-Elvira et al. 2018). Despite being described for *K. lactis* (Hernández-Elvira et al. 2018), the UPR has not yet been described in *K. marxianus*. Venturing in this endeavor could be an opportunity to understand how *K. marxianus* adapts to ethanol stress and could help to better understand its low tolerance to ethanol.

Oxidative stress caused by ethanol and its associated response

In eukaryotes, one of the primary responses is the conversion of ethanol to acetaldehyde through the activity of the enzyme alcohol dehydrogenase (ADH) (de Smidt et al. 2012). However, acetaldehyde is a notorious producer of ROS in the mitochondria, resulting in oxidative stress and further damage to the cell (Novitskiy et al. 2006; Yan and Zhao 2020). As ethanol disturbs mitochondrial function, this in turn affects flux in the central metabolic pathways. The cell must produce additional redox

and energy cofactors to compensate for the reduction in proton motive force in the mitochondria and combat ROS (Yang et al. 2012).

In *S. cerevisiae*, the increased demand for NAD(P)H increases the flux through the pentose phosphate pathway (PPP), resulting in accumulation of 6-phosphogluconate and of all intermediated of the non-oxidative branch of the PPP (Celton et al. 2012). *K. marxianus* counters oxidative stress and responds to the increased demand of reduced cofactors by up-regulating a number of genes encoding enzymes related to NAD(P)H generation, which is necessary to detoxify the cell of ROS. One such enzyme is enoate reductase 1, which is known in *S. cerevisiae* to catalyze the production of 2-butenoate and reduced NADH (Trotter et al. 2006; Odat et al. 2007). This up-regulation of enoate reductase 1 was detected in a proteomics analysis performed by Alvim et al. (2019). They also observed that the Adh1p and Adh2p enzymes were more abundant during stress. Surprisingly, the NAD(P)-dependent D-xylose reductase was also up-regulated, even though no xylose was present in the culture medium. Other omics studies also detected the up-regulation of these proteins (Schabort et al. 2016; Wang et al. 2018). Diniz et al (2017) found that the gene *ZWF1*, which encodes glucose-6-phosphate-1-dehydrogenase, is up-regulated. This suggests that the redirection of flux towards PPP for NADPH regeneration is also present in *K. marxianus*. A more direct countermeasure against ROS was further demonstrated by Wang et al. (2018), which detected an up-regulation of the superoxide dismutase (SODs) genes *SOD1* and *SOD2*, as well as two isoforms of catalase, *CTA1* and *CTT1*, which are all essential to convert ROS to inert forms.

An important cellular response to oxidative stress caused by ethanol is autophagy. It is a process in which damaged or unnecessary cellular components such as macromolecules, membranes and organelles are degraded inside vacuoles (Reggiori and Klionsky 2013). Autophagy can be triggered by natural cellular processes that are necessary for sustaining homeostasis or from stressful conditions. In yeasts, it has been shown that the ROS generated during ethanol stress can trigger non-selective macroautophagy (Jing et al. 2018). Macroautophagy employs a double-membrane structure called the autophagosome, whose function is regulated by proteins coded by the *ATG* gene family (Mizushima et al. 2011). In *S. cerevisiae*, the production of ROS and the expression of *ATG8* and *ATG1* genes were evaluated during ethanol stress by Jing et al. (2018). They reported that the expression of these

genes increased with rising levels of ROS which in turn activated autophagy. In a study performed by Gao et al. (2015), the strain *K. marxianus* Y179 was used to produce ethanol from inulin as a carbon source. The authors performed a transcriptomics analysis and noticed an up-regulation of the genes *ATG3*, *ATG7* and *ATG8*, which are known to contribute to autophagosome formation (Reggiori and Klionsky 2013). Fu et al. (2019) also performed a transcriptomics analysis, using the strain *K. marxianus* DMKU3-1042 after an arrest during high-temperature ethanol fermentation. They detected an up-regulation of the genes *VAC8*, *ARF1*, *ATG3*, *ATG8* and *PRB1*, all of which are necessary for assembly and function of the autophagosome. This suggests that non-selective macroautophagy plays an important role in helping the cell deal with damage caused by ethanol-induced ROS. By recycling damaged macromolecules and damaged organelles, the deleterious effects of dysfunctional cell components can be avoided, thus assisting in cell survival.

Metabolite concentration changes in response to ethanol stress

Aside from the metabolites mentioned previously in this review, that is, ergosterol, unsaturated fatty acids, trehalose and intermediates of the non-oxidative branch of the PPP, other metabolites undergo changes upon ethanol stress. Such metabolites include carbohydrates, organic acids and amino acids. In *S. cerevisiae*, the redox imbalance caused by the conversion of ethanol to acetaldehyde leads to an accumulation of glycerol, which could work as an alternative way to regenerate NAD^+ (Vriesekoop et al. 2009). Increased glycerol production is also related to osmo-protection, as it contributes to maintain osmotic pressure via serving as an osmolyte (Jhariya et al. 2021) An up-regulation of the genes *GPD1*, *GPD2*, *GUT1*, *GUT2*, *HOR2*, *RHR2*, which are involved with the glycerol biosynthesis, were detected after submitting yeast cells to a 6% (v/v) ethanol shock in batch growth conditions (Li et al. 2009). This finding agrees with the observed increase in glycerol concentration as measured by HPLC by Vriesekoop et al. (2009). For *K. marxianus*, Mo et al. (2019) did not observe changes in gene expression of glycerol biosynthetic genes. In agreement with these results, Alvim et al. (2019) did not detect an increase in glycerol concentration in their metabolomics analysis. In this experiment, the strain CCT 7735 was cultivated in batch culture conditions added with 6% (v/v) ethanol. As aforementioned, the regeneration of NAD^+ could come from the up-regulation of a

number of genes encoding enzymes related to NAD(P)H generation and their respective reactions, such as the reduction of NAD^+ to NADH through the production of 2-butenoate (Trotter et al. 2006; Odat et al. 2007). For osmo-protection, Mo et al. (2019) argues that another mechanism could be in effect, although this remains to be studied in more depth.

The disturbance of the central metabolic pathways and the increased demand for NAD(P)H leads to an accumulation and depletion of a series of organic acids, some of which are intermediates in these pathways. The increased production of acetic acid is one example. The oxidation of acetaldehyde yields one molecule of acetate and one of NADH (Curiel et al. 2016). This increase in acetate concentration was detected in a study by Stanley et al. (2010), using ethanol-tolerant mutant *S. cerevisiae* strains. Other organic acids such as the intermediate metabolites of glycolysis and the TCA cycle are also affected by ethanol stress and had their concentrations altered. A metabolomics survey by Ming et al. (2019) using *S. cerevisiae* S288c have shown that oxalate, D-lactate, citrate, succinate and 4-hydroxybutyrate concentrations increased, while the concentrations of 3-hydroxybutyrate, fumarate and pyruvate decreased. In this regard, for *K. marxianus* the increase in acetate concentration was also reported by Fu et al. (2019), in their analysis using the strain *K. marxianus* DMKU3-1042 after an arrest during high-temperature ethanol fermentation. The authors also correlated this result with NAD(P)H regeneration. However, when cultivating *K. marxianus* at 30°C, they did not detect a change in pyruvate concentrations. This was also not detected by the metabolomics analysis of Alvim et al. (2019). However, Fu et al. (2019) observed a decreased concentration of several other organic acids, such as glyceraldehyde, D-glycerate, D-gluconate, L-glucono-1,4-lactone, malate and oxaloacetate.

Regarding changes in amino acids, Ming et al. (2019) reported an increase of alanine and methionine, and a decrease in proline, isoleucine, threonine, aspartate and glutamate in *S. cerevisiae*. In *K. marxianus* CCT 7735, there was a decrease in arginine, proline, glutamate, phenylalanine and tryptophan concentrations after ethanol exposure (Alvim et al. 2019). However, at a later time of ethanol exposure, these authors observed that the concentrations of these amino acids were higher in stressed cells than in unstressed cells, suggesting that they play a role in coping with the stress condition. Consistent with this, proline and tryptophan have a protective

effect against ethanol stress (Jhariya et al. 2021). Nevertheless, it is noteworthy that metabolite changes could be also related to a leakage of amino acids through a damaged cell membrane under ethanol stress conditions (Silveira et al. 2020; Jhariya et al. 2021).

Selection of *K. marxianus* strains with enhanced tolerance to ethanol by metabolic engineering strategies

Adaptive laboratory Evolution (ALE) and Directed Evolution (DE) are valuable tools of metabolic engineering that can be used to select microbial strains with desirable features in bioprocesses. Recently, both ALE and ED were used to select *K. marxianus* ethanol-tolerant strains.

ALE has also been used to select microbial cell factories tolerant to stress conditions found in bioprocesses, as well as with improved nutrient uptake and productivity of metabolites (Portnoy et al. 2011; Dragosits and Mattanovich 2013; Pennisi 2013; LaCroix et al. 2017). In ALE experiments, cells are grown under defined conditions for a long period of time (Portnoy et al. 2011; Dragosits and Mattanovich 2013; LaCroix et al. 2017), and randomized non-directed mutations accumulate in genomes of microorganisms, leading to the changes desirable for using them in bioprocesses (Barrick et al. 2009; Portnoy et al. 2011; Dragosits and Mattanovich 2013; LaCroix et al. 2017; Tokuyama et al. 2018).

The *K. marxianus* FIM1 strain was subjected to ALE in a culture medium containing 6% (v/v) ethanol in batch experiments (Mo et al. 2019). Ethanol-tolerant strains were selected after 450 generations (100-days evolution). Importantly, the ethanol tolerant strain also presented tolerance to other stress conditions, that is, temperature, osmotic, and oxidative stresses. In addition, some genes encoding proteins related to pathways such as ethanol consumption, membrane lipid biosynthesis, protein folding, response to osmotic pressure and oxidative stress are overexpressed even in the absence of ethanol. It is noteworthy that there was an increase in expression of genes encoding enzymes that participate in pathways involved with both phosphoglyceride and ergosterol biosynthesis at high ethanol concentration. This indicates that alterations in cell membrane contributed to enhance the ethanol tolerance in *K. marxianus*.

Another strain, *K. marxianus* CCT 7735, was also submitted to ALE aiming the selection of strains with enhanced tolerance to high ethanol concentrations. Four ethanol-tolerant strains (ETS1, ETS2, ETS3 and ETS4) were obtained by serial passages across 300 generations under ethanol 4% (v/v), a period in which there was a significant increase (above 50%) in the specific growth rate (Silveira et al. 2020). The *K. marxianus* ETS4 was selected for also displaying a specific ethanol production rate higher than the parental strain. Compared to the parental strain, the ETS4 accumulated amino acids and metabolites of TCA cycle in response to ethanol. In addition, the ethanol-tolerant strain displayed higher contents of fatty acids and ergosterol than the parental strain. These changes contributed to maintaining the structural integrity of the cell membrane as metabolite leakage was lower in the ethanol-tolerant strain than in the parental. Consistent with these findings, genomic analysis identified mutations in genes associated with the lipid metabolism and regulation of ergosterol biosynthesis.

ALE was also applied for the *K. marxianus* MTCC1389 strain aiming its adaptation in culture media containing a high lactose concentration (200 $g.L^{-1}$ lactose) (Saini et al. 2017). Interestingly, the evolved strains also display an improved tolerance to high temperatures and ethanol concentrations.

Besides ALE, DE also has been developed to select microbial strains tolerant to stressful conditions (Fong et al. 2005; Lee et al. 2005; Steensels et al. 2014; Li et al. 2018; Reed et al. 2019). DE is a valuable technique to generate a large gene variant library. Upon library construction, a screening step is required to select the gene variants that display suitable features; therefore, it has been widely used for protein engineering (Packer and Liu 2015). Li et al. (2018) used error-prone PCR to construct a gene variant library of the gene encoding the *K. marxianus* TATA-binding protein Spt15, one of the components of the general factor RNA polymerase II transcription factor D. After the screening, two mutant ethanol-tolerant strains were selected. The strain M2 also stood out for displaying an increased ethanol productivity. In this strain, whose Spt15 protein has a single amino acid substitution, the expression of a hundred genes was altered. Genes encoding proteins related to amino acid transport, long-chain fatty acid biosynthesis and MAPK signaling pathway were up-regulated, whilst genes encoding proteins associated with ribosome biogenesis, translation and protein synthesis were down-regulated.

Conclusions and future perspectives

K. marxianus has great potential for industrial applications such as second generation ethanol production, but its low tolerance to ethanol hinders its broader applicability. The effects of ethanol inside the cell are concentrated primarily in cell membranes, proteins and metabolism. The ethanol stress response cross-talks with the oxidative stress response, as many mechanisms and pathways are shared between them. Ethanol is even capable of triggering an increase in ROS, further exacerbating the oxidative stress response. In yeasts, the cell response pathways include most notably changes in lipid metabolism, accumulation of metabolites such as trehalose, PQC-associated responses, UPR, ROS detoxification and non-selective macroautophagy. Although recent studies have provided some cues regarding ethanol stress responses in *K. marxianus*, it is still not clear why it is less tolerant to ethanol than other yeasts. Future studies concerning the UPR, which is still not described in *K. marxianus*, could help explain why this happens. The solution to this problem might hold the key to rationally engineer *K. marxianus* strains to achieve tolerance to high levels of ethanol. Still, many endeavors are already taking place, especially with undirected metabolic engineering tools such as ALE and DE. Thus, understanding the ethanol stress response in *K. marxianus* is important for enhancing its capabilities and industrial applications.

Many of the studies that uncovered these responses were based on omics technologies, especially gene expression and metabolite analysis. The available data could be useful for systems biology approaches such as constraint-based modeling and gene regulatory networks. A genome-scale reconstruction of *K. marxianus* is already available, and it was used to model thermotolerance (Marcišauskas et al. 2019). Similarly, it could be used to model ethanol fermentation and stress. Inferences of regulatory networks using transcriptomics data could also uncover many mechanisms and aid bioengineering strategies, as has been done with *S. cerevisiae* (Liu et al. 2019). Comparative genomics analysis could also be employed to ALE strains and be used to compare the different mutations and changes in phenotype among different evolved strains. Further, the use of machine learning models and protein engineering strategies can be useful to make more stable proteins (Rouvinen et al. 2021).

Declarations

Funding

This work was supported by the Brazilian National Council for Scientific and Technological Development (CNPq); the Foundation for Research Support of the State of Minas Gerais (FAPEMIG); and the Coordination for the Improvement of Higher Education Personnel (CAPES, Finance Code 001)

References

Adams CJ, Kopp MC, Larburu N, Nowak PR, Ali MMU (2019) Structure and molecular mechanism of ER stress signaling by the unfolded protein response signal activator IRE1. Front Mol Biosci 6:1–12 . https://doi.org/10.3389/fmolb.2019.00011

Alvim MCT, Vital CE, Barros E, Vieira NM, da Silveira FA, Balbino TR, Diniz RHS, Brito AF, Bazzolli DMS, de Oliveira Ramos HJ, da Silveira WB (2019) Ethanol stress responses of *Kluyveromyces marxianus* CCT 7735 revealed by proteomic and metabolomic analyses. Antonie van Leeuwenhoek, Int J Gen Mol Microbiol 112:827–845 . https://doi.org/10.1007/s10482-018-01214-y

Bajpai P, Margaritis A (1982) Ethanol inhibition kinetics of *Kluyveromyces marxianus* grown on Jerusalem artichoke juice. Appl Environ Microbiol 44:1325–1329 . https://doi.org/10.1128/aem.44.6.1325-1329.1982

Balat M, Balat H (2009) Recent trends in global production and utilization of bio-ethanol fuel. Appl Energy 86:2273–2282 . https://doi.org/10.1016/j.apenergy.2009.03.015

Bandara A, Fraser S, Chambers PJ, Stanley GA (2009) Trehalose promotes the survival of *Saccharomyces cerevisiae* during lethal ethanol stress, but does not influence growth under sublethal ethanol stress. FEMS Yeast Res 9:1208–1216 . https://doi.org/10.1111/j.1567-1364.2009.00569.x

Barrick JE, Yu DS, Yoon SH, Jeong H, Oh TK, Schneider D, Lenski RE, Kim JF (2009) Genome evolution and adaptation in a long-term experiment with *Escherichia coli*. Nature 461:1243–1247 . https://doi.org/10.1038/nature08480

Buck M (1998) Trifluoroethanol and colleagues: cosolvents come of age. Recent studies with peptides and proteins. Q Rev Biophys 31:297–355 . https://doi.org/10.1017/S003358359800345X

Buenrostro-Figueroa J, Tafolla-Arellano JC, Flores-Gallegos AC, Rodríguez-Herrera

R, De la garza-Toledo H, Aguilar CN (2018) Native yeasts for alternative utilization of overripe mango pulp for ethanol production. Rev Argent Microbiol 50:173–177 . https://doi.org/10.1016/j.ram.2016.04.010

Campos BB, Diniz RHS, Da Silveira FA, Ribeiro Júnior JI, Fietto LG, MacHado JC, Da Silveira WB (2019) Elephant grass (*Pennisetum purpureum* Schumach) is a promising feedstock for ethanol production by the thermotolerant yeast *Kluyveromyces marxianus* CCT 7735. Brazilian J Chem Eng 36:43–49 . https://doi.org/10.1590/0104-6632.20190361s20170263

Caspeta L, Chen Y, Ghiaci P, Feizi A, Baskov S, Hallström BM, Petranovic D, Nielsen J (2014) Altered sterol composition renders yeast thermotolerant. Science (80-) 346:75–78 . https://doi.org/10.1126/SCIENCE.1258137

Celton M, Sanchez I, Goelzer A, Fromion V, Camarasa C, Dequin S (2012) A comparative transcriptomic, fluxomic and metabolomic analysis of the response of *Saccharomyces cerevisiae* to increases in NADPH oxidation. BMC Genomics 13: . https://doi.org/10.1186/1471-2164-13-317

Charoenbhakdi S, Dokpikul T, Burphan T, Techo T, Auesukaree C (2016) Vacuolar H+-ATPase protects *Saccharomyces cerevisiae* cells against ethanol induced oxidative and cell wall stresses. Appl Environ Microbiol 82:3121–3130 . https://doi.org/10.1128/AEM.00376-16

Chen B, Retzlaff M, Roos T, Frydman J (2011) Cellular Strategies of Protein Quality Control. Cold Spring Harb Perspect Biol 3:a004374, https://doi.org/10.1101/cshperspect.a004374

Costa DA, de Souza CJA, Costa PA, Rodrigues MQRB, dos Santos AF, Lopes MR, Genier HLA, Silveira WB, Fietto LG (2014) Physiological characterization of thermotolerant yeast for cellulosic ethanol production. Appl Microbiol Biotechnol 98:3829–40 . https://doi.org/10.1007/s00253-014-5580-3

Cray JA, Russell JT, Timson DJ, Singhal RS, Hallsworth JE (2013) A universal measure of chaotropicity and kosmotropicity. Environ Microbiol 15:287–296 . https://doi.org/10.1111/1462-2920.12018

Curiel JA, Salvadó Z, Tronchoni J, Morales P, Rodrigues AJ, Quirós M, Gonzalez (2016) Identification of target genes to control acetate yield during aerobic fermentation with *Saccharomyces cerevisiae*. Microb Cell Fact 15:156, https://doi.org/10.1186/s12934-016-0555-y

Dasgupta D, Suman SK, Pandey D, Ghosh D, Khan R, Agrawal D, Jain RK, Vadde VT, Adhikari DK (2013) Design and optimization of ethanol production from bagasse pith hydrolysate by a thermotolerant yeast *Kluyveromyces* sp. IIPE453 using response surface methodology. Springer plus 2:1–10 . https://doi.org/10.1186/2193-1801-2-159

de Smidt O, du Preez JC, Albertyn J (2012) Molecular and physiological aspects of alcohol dehydrogenases in the ethanol metabolism of *Saccharomyces cerevisiae*. FEMS Yeast Res 12:33–47 . https://doi.org/10.1111/j.1567-1364.2011.00760.x

de Souza CJA, Costa DA, Rodrigues MQRB, dos Santos AF, Lopes MR, Abrantes ABP, dos Santos Costa P, Silveira WB, Passos FML, Fietto LG (2012) The influence of presaccharification, fermentation temperature and yeast strain on ethanol production from sugarcane bagasse. Bioresour Technol 109:63–69 . https://doi.org/10.1016/j.biortech.2012.01.024

Deparis Q, Claes A, Foulquié-Moreno MR, Thevelein JM (2017) Engineering tolerance to industrially relevant stress factors in yeast cell factories. FEMS Yeast Res 17:1–35 . https://doi.org/10.1093/femsyr/fox036

Dequin S, Casaregola S (2011) The genomes of fermentative *Saccharomyces*. Comptes Rendus - Biol 334:687–693 . https://doi.org/10.1016/j.crvi.2011.05.019

Devanand T, Krishnaswamy S, Vemparala S (2019) Interdigitation of Lipids Induced by Membrane–Active Proteins. J Membr Biol 252:331–342 . https://doi.org/10.1007/s00232-019-00072-7

Ding J, Huang X, Zhang L, Zhao N, Yang D, Zhang K (2009) Tolerance and stress response to ethanol in the yeast *Saccharomyces cerevisiae*. Appl Microbiol Biotechnol 85:253–263 . https://doi.org/10.1007/s00253-009-2223-1

Diniz RHS, Rodrigues MQRB, Fietto LG, Passos FML, Silveira WB (2013) Optimizing and validating the production of ethanol from cheese whey permeate by *Kluyveromyces marxianus* UFV-3. Biocatal Agric Biotechnol 3:111–117 . https://doi.org/10.1016/j.bcab.2013.09.002

Diniz RHS, Villada JC, Alvim MCT, Vidigal PMP, Vieira NM, Lamas-Maceiras M, Cerdán ME, González-Siso MI, Lahtvee PJ, Silveira WB da (2017) Transcriptome analysis of the thermotolerant yeast *Kluyveromyces marxianus* CCT 7735 under ethanol stress. Appl Microbiol Biotechnol 101:6969–6980 . https://doi.org/10.1007/s00253-017-8432-0

Dos Santos VC, Bragança CRS, Passos FJV, Passos FML (2013) Kinetics of growth and ethanol formation from a mix of glucose/xylose substrate by *Kluyveromyces marxianus* UFV-3. Antonie van Leeuwenhoek, Int J Gen Mol Microbiol 103:153–161 . https://doi.org/10.1007/s10482-012-9794-z

Dragosits M, Mattanovich D (2013) Adaptive laboratory evolution – principles and applications for biotechnology. Microb Cell Fact 12 VN-r:64 . https://doi.org/10.1186/1475-2859-12-64

Eardley J, Timson DJ (2020) Yeast Cellular Stress: Impacts on Bioethanol Production. Fermentation 6:109 . https://doi.org/10.3390/fermentation6040109

Ferreira PG, da Silveira FA, dos Santos RCV, Genier HLA, Diniz RHS, Ribeiro JI, Fietto LG, Passos FML, da Silveira WB (2015) Optimizing ethanol production by thermotolerant *Kluyveromyces marxianus* CCT 7735 in a mixture of sugarcane bagasse and ricotta whey. Food Sci Biotechnol 24:1421–1427 . https://doi.org/10.1007/s10068-015-0182-0

Fong SS, Burgard AP, Herring CD, Knight EM, Blattner FR, Maranas CD, Palsson BO (2005) In silico design and adaptive evolution of *Escherichia coli* for production of lactic acid. Biotechnol Bioeng 91:643–648 . https://doi.org/10.1002/bit.20542

Fu X, Li P, Zhang L, Li S (2019) Understanding the stress responses of *Kluyveromyces marxianus* after an arrest during high-temperature ethanol fermentation based on integration of RNA-Seq and metabolite data. Appl Microbiol Biotechnol 103:2715–2729 . https://doi.org/10.1007/s00253-019-09637-x

Gao J, Yuan W, Li Y, Xiang R, Hou S, Zhong S, Bai F (2015) Transcriptional analysis of *Kluyveromyces marxianus* for ethanol production from inulin using consolidated bioprocessing technology. Biotechnol Biofuels 8:1–17 . https://doi.org/10.1186/s13068-015-0295-y

Gomez M, Pérez-Gallardo R V., Sánchez LA, Díaz-Pérez AL, Cortés-Rojo C, Carmen VM, Saavedra-Molina A, Lara-Romero J, Jiménez-Sandoval S, Rodríguez F, Rodríguez-Zavala JS, Campos-García J (2014) Malfunctioning of the iron-sulfur cluster assembly machinery in saccharomyces cerevisiae produces oxidative stress via an iron-dependent mechanism, causing dysfunction in respiratory complexes. PLoS One 9: . https://doi.org/10.1371/journal.pone.0111585

Goshima T, Negi K, Tsuji M, Inoue H, Yano S, Hoshino T, Matsushika A (2013) Ethanol fermentation from xylose by metabolically engineered strains of *Kluyveromyces*

marxianus. J Biosci Bioeng 116:551–554 . https://doi.org/10.1016/j.jbiosc.2013.05.010

Güneşer O, Karagül-Yüceer Y, Wilkowska A, Kregiel D (2016) Volatile metabolites produced from agro-industrial wastes by Na-alginate entrapped *Kluyveromyces marxianus*. Brazilian J Microbiol 47:965–972 . https://doi.org/10.1016/j.bjm.2016.07.018

Halbleib K, Pesek K, Covino R, Hofbauer HF, Wunnicke D, Hänelt I, Hummer G, Ernst R (2017) Activation of the Unfolded Protein Response by Lipid Bilayer Stress. Mol Cell 67:673-684.e8 . https://doi.org/10.1016/j.molcel.2017.06.012

Hernández-Elvira M, Torres-Quiroz F, Escamilla-Ayala A, Domínguez-Martin E, Escalante R, Kawasaki L, Ongay-Larios L, Coria R (2018) The Unfolded Protein Response Pathway in the Yeast *Kluyveromyces lactis*. A Comparative View among Yeast Species. Cells 2018, Vol 7, Page 106 7:106 . https://doi.org/10.3390/CELLS7080106

Hesham AEL, Wambui V, Ogola J.O. H, Maina JM (2014) Phylogenetic analysis of isolated biofuel yeasts based on 5.8S-ITS rDNA and D1/D2 26S rDNA sequences. J Genet Eng Biotechnol 12:37–43 . https://doi.org/10.1016/j.jgeb.2014.01.001

Ho C, Grousl T, Shatz O, Jawed A, Ruger-Herreros C, Semmelink M, Zahn R, Richter K, Bukau B, Mogk A (2019) Cellular sequestrases maintain basal Hsp70 capacity ensuring balanced proteostasis. Nat Commun 10, 4851, https://doi.org/10.1038/s41467-019-12868-1

IEA IEA, Cooper G, McCaherty J, Huschitt E, Schwarck R, Wilson C (2021) 2021 Ethanol Industry Outlook. Renew fuels Assoc 1–40

IEA IEA, Koehler N, Mccaherty J, Wilson C, Cooper G, Schwarck R, Kemmet N, Baker R, Mcafee E, Drook R, Markham S, Sitzmann C, Friedberg J, Ricketts M, Christensen S, Boyle P, Harder S, Keiser K, Woodside C, Roe S, Wilson C (2020) Focus forward. 2020 RFA's Ethanol Industry Outlook. Renew Fuels Assoc Ellisvile, MO, USA

Jhariya U, Dafale NA, Srivastava S, Bhende RS, Kapley A, Purohit HJ (2021) Understanding Ethanol Tolerance Mechanism in *Saccharomyces cerevisiae* to Enhance the Bioethanol Production: Current and Future Prospects. Bioenergy Res 14:670–688 . https://doi.org/10.1007/s12155-020-10228-2

Jing H, Liu H, Zhang L, Gao J, Song H, Tan X (2018) Ethanol induces autophagy

regulated by mitochondrial ROS in *Saccharomyces cerevisiae*. J Microbiol Biotechnol 28:1982–1991 . https://doi.org/10.4014/jmb.1806.06014

Jung YR, Park JM, Heo SY, Hong WK, Lee SM, Oh BR, Park SM, Seo JW, Kim CH (2015) Cellulolytic enzymes produced by a newly isolated soil fungus *Penicillium* sp. TG2 with potential for use in cellulosic ethanol production. Renew Energy 76:66–71 . https://doi.org/10.1016/j.renene.2014.10.064

Kádár Z, Szengyel Z, Réczey K (2004) Simultaneous saccharification and fermentation (SSF) of industrial wastes for the production of ethanol. Ind Crops Prod 20:103–110 . https://doi.org/10.1016/j.indcrop.2003.12.015

Koutinas M, Menelaou M, Nicolaou EN (2014) Development of a hybrid fermentation-enzymatic bioprocess for the production of ethyl lactate from dairy waste. Bioresour Technol 165:343–349 . https://doi.org/10.1016/j.biortech.2014.03.053

LaCroix RA, Palsson BO, Fiest AM (2017) A model for Designing Adaptive Laboratory Evolution Experiments. Appl Enviromental Microbiol 83:1–14 . https://doi.org/https://doi.org/ 10.1128/AEM.03115-16.

Lee S, Lee D, Kim T, Kim B (2005) Metabolic engineering of *Escherichia coli* for enhanced production of succinic acid, based on genome comparison and *in silico* gene knockout simulation. Appl Enviromental Microbiol 71:7880–7887 . https://doi.org/10.1128/AEM.71.12.7880

Li P, Fu X, Chen M, Zhang L, Li S (2019) Proteomic profiling and integrated analysis with transcriptomic data bring new insights in the stress responses of *Kluyveromyces marxianus* after an arrest during high-temperature ethanol fermentation. Biotechnol Biofuels 12:1–13 . https://doi.org/10.1186/s13068-019-1390-2

Li P, Fu X, Li S, Zhang L (2018) Engineering TATA-binding protein Spt15 to improve ethanol tolerance and production in *Kluyveromyces marxianus*. Biotechnol Biofuels 11:1–13 . https://doi.org/10.1186/s13068-018-1206-9

Limayem A, Ricke SC (2012) Lignocellulosic biomass for bioethanol production: Current perspectives, potential issues and future prospects. Prog Energy Combust Sci 38:449–467 . https://doi.org/10.1016/j.pecs.2012.03.002

Limtong S, Sringiew C, Yongmanitchai W (2007) Production of fuel ethanol at high temperature from sugar cane juice by a newly isolated *Kluyveromyces marxianus*. Bioresour Technol 98:3367–3374 . https://doi.org/10.1016/j.biortech.2006.10.044

Ma M, Liu ZL (2010) Mechanisms of ethanol tolerance in *Saccharomyces cerevisiae*. Appl Microbiol Biotechnol 87:829–845 . https://doi.org/10.1007/s00253-010-2594-3

Madeira-Jr JV, Gombert AK (2018) Towards high-temperature fuel ethanol production using *Kluyveromyces marxianus*: On the search for plug-in strains for the Brazilian sugarcane-based biorefinery. Biomass and Bioenergy 119:217–228 . https://doi.org/10.1016/j.biombioe.2018.09.010

Martin SR, Esposito V, Rios PDL, Pastore A, Temussi PA (2008) Cold Denaturation of Yeast Frataxin Offers the Clue to Understand the Effect of Alcohols on Protein Stability. J Am Chem Soc 130:9963–9970 . https://doi.org/10.1021/JA803280E

Mehmood N, Alayoubi R, Husson E, Jacquard C, Büchs J, Sarazin C, Gosselin I (2018) *Kluyveromyces marxianus*, an attractive yeast for ethanolic fermentation in the presence of imidazolium ionic liquids. Int J Mol Sci 19: . https://doi.org/10.3390/ijms19030887

Ming M, Wang X, Lian L, Zhang H, Gao W, Zhu B, Low D (2019) Metabolic responses of *Saccharomyces cerevisiae* to ethanol stress using gas chromatography-mass spectrometry. Mol Omics 15:216, https://doi.org/ 10.1039/c9mo00055k

Mizushima N, Yoshimori T, Ohsumi Y (2011) The role of atg proteins in autophagosome formation. Annu Rev Cell Dev Biol 27:107–132 . https://doi.org/10.1146/annurev-cellbio-092910-154005

Mo W, Wang M, Zhan R, Yu Y, He Y, Lu H (2019) *Kluyveromyces marxianus* developing ethanol tolerance during adaptive evolution with significant improvements of multiple pathways. Biotechnol Biofuels 12:1–15 . https://doi.org/10.1186/s13068-019-1393-z

Moncan M, Mnich K, Blomme A, Almanza A, Samali A, Gorman AM (2021) Regulation of lipid metabolism by the unfolded protein response. J Cell Mol Med 25:1359–1370 . https://doi.org/10.1111/jcmm.16255

Moreno AD, Ibarra D, Ballesteros I, González A, Ballesteros M (2013) Comparing cell viability and ethanol fermentation of the thermotolerant yeast *Kluyveromyces marxianus* and *Saccharomyces cerevisiae* on steam-exploded biomass treated with laccase. Bioresour Technol 135:239–245 . https://doi.org/10.1016/j.biortech.2012.11.095

Naik SN, Goud V V., Rout PK, Dalai AK (2010) Production of first and second generation biofuels: A comprehensive review. Renew Sustain Energy Rev 14:578–

597 . https://doi.org/10.1016/j.rser.2009.10.003

Nakamura K, Shinomiya N, Orikasa Y, Oda Y (2012) Efficient production of ethanol from saccharified crops mixed with cheese whey by the flex yeast *Kluyveromyces marxianus* KD-15. Food Sci Technol Res 18:235–242 . https://doi.org/10.3136/fstr.18.235

Navarro-Tapia E, Pérez-Torrado R, Querol A (2017) Ethanol effects involve non-canonical unfolded protein response activation in yeast cells. Front Microbiol 8:1–12 . https://doi.org/10.3389/fmicb.2017.00383

Nikolaidis A, Andreadis M, Moschakis T (2017) Effect of heat, pH, ultrasonication and ethanol on the denaturation of whey protein isolate using a newly developed approach in the analysis of difference-UV spectra. Food Chem 232:425–433 . https://doi.org/10.1016/j.foodchem.2017.04.022

Novitskiy G, Traore K, Wang L, Trush MA, Mezey E (2006) Effects of ethanol and acetaldehyde on reactive oxygen species production in rat hepatic stellate cells. Alcohol Clin Exp Res 30:1429–1435 . https://doi.org/10.1111/j.1530-0277.2006.00171.x

Odat O, Matta S, Khalil H, Kampranis SC, Pfau R, Tsichlis PN, Makris AM (2007) Old Yellow Enzymes, Highly Homologous FMN Oxidoreductases with Modulating Roles in Oxidative Stress and Programmed Cell Death in Yeast *. J Biol Chem 282:36010–36023 . https://doi.org/10.1074/JBC.M704058200

Ortiz-Merino RA, Varela JA, Coughlan AY, Hoshida H, da Silveira WB, Wilde C, Kuijpers NGA, Geertman JM, Wolfe KH, Morrissey JP (2018) Ploidy variation in *Kluyveromyces marxianus* separates dairy and non-dairy isolates. Front Genet 9:1–16 . https://doi.org/10.3389/fgene.2018.00094

Packer MS, Liu DR (2015) Methods for the directed evolution of proteins. Nat Rev Genet 16:379–394 . https://doi.org/10.1038/nrg3927

Pennisi E (2013) The man who bottled evolution. Science (80-) 342:790–793

Pereira FB, Guimarães PMR, Teixeira JA, Domingues L (2011) Robust industrial *Saccharomyces cerevisiae* strains for very high gravity bio-ethanol fermentations. J Biosci Bioeng 112:130–136 . https://doi.org/10.1016/j.jbiosc.2011.03.022

Pilap W, Thanonkeo S, Klanrit P, Thanonkeo P (2018) The potential of the newly isolated thermotolerant *Kluyveromyces marxianus* for high-temperature ethanol production using sweet sorghum juice. 3 Biotech 8: .

https://doi.org/10.1007/s13205-018-1161-y

Piškur J, Rozpedowska E, Polakova S, Merico A, Compagno C (2006) How did *Saccharomyces* evolve to become a good brewer? Trends Genet 22:183–186 . https://doi.org/10.1016/j.tig.2006.02.002

Portnoy VA, Bezdan D, Zengler K (2011) Adaptive laboratory evolution—harnessing the power of biology for metabolic engineering. Curr Opin Biotechnol 22:590–594 . https://doi.org/10.1016/j.copbio.2011.03.007

Read A, Schröder M (2021) The Unfolded Protein Response: An Overview. Biol 2021, Vol 10, Page 384 10:384 . https://doi.org/10.3390/BIOLOGY10050384

Reed KB, Wagner JM, d'Oelsnitz S, Wiggers JM, Alper HS (2019) Improving ionic liquid tolerance in *Saccharomyces cerevisiae* through heterologous expression and directed evolution of an ILT1 homolog from *Yarrowia lipolytica*. J Ind Microbiol Biotechnol 46:1715–1724 . https://doi.org/10.1007/s10295-019-02228-9

Reggiori F, Klionsky DJ (2013) Autophagic processes in yeast: Mechanism, machinery and regulation. Genetics 194:341–361 . https://doi.org/10.1534/genetics.112.149013

Rugthaworn P, Murata Y, Machida M, Apiwatanapiwat W, Hirooka A, Thanapase W, Dangjarean H, Ushiwaka S, Morimitsu K, Kosugi A, Arai T, Vaithanomsat P (2014) Growth inhibition of thermotolerant yeast, *Kluyveromyces marxianus*, in hydrolysates from cassava pulp. Appl Biochem Biotechnol 173:1197–1208 . https://doi.org/10.1007/s12010-014-0906-2

Saini P, Beniwal A, Kokkiligadda A, Vij S (2018) Response and tolerance of yeast to changing environmental stress during ethanol fermentation. Process Biochem 72:1–12 . https://doi.org/10.1016/j.procbio.2018.07.001

Saini P, Beniwal A, Kokkiligadda A, Vij S (2017) Evolutionary adaptation of *Kluyveromyces marxianus* strain for efficient conversion of whey lactose to bioethanol. Process Biochem. https://doi.org/10.1016/j.procbio.2017.07.013

Sandoval-Nuñez D, Arellano-Plaza M, Gschaedler A, Arrizon J, Amaya-Delgado L (2018) A comparative study of lignocellulosic ethanol productivities by *Kluyveromyces marxianus* and *Saccharomyces cerevisiae*. Clean Technol Environ Policy 20:1491–1499 . https://doi.org/10.1007/s10098-017-1470-6

Sansonetti S, Curcio S, Calabrò V, Iorio G (2009) Bio-ethanol production by fermentation of ricotta cheese whey as an effective alternative non-vegetable

source. Biomass and Bioenergy 33:1687–1692 . https://doi.org/10.1016/j.biombioe.2009.09.002

Saratale GD, Saratale RG, Ghodake GS, Jiang YY, Chang JS, Shin HS, Kumar G (2017) Solid state fermentative lignocellulolytic enzymes production, characterization and its application in the saccharification of rice waste biomass for ethanol production: An integrated biotechnological approach. J Taiwan Inst Chem Eng 76:51–58 . https://doi.org/10.1016/j.jtice.2017.03.027

Schabort DTWP, Letebele PK, Steyn L, Kilian SG, Preez JC du (2016) Differential RNA-seq, Multi-Network Analysis and Metabolic Regulation Analysis of *Kluyveromyces marxianus* Reveals a Compartmentalised Response to Xylose. PLoS One 11:e0156242 . https://doi.org/10.1371/JOURNAL.PONE.0156242

Silveira FA, de Oliveira Soares DL, Bang KW, Balbino TR, de Moura Ferreira MA, Diniz RHS, de Lima LA, Brandão MM, Villas-Bôas SG, da Silveira WB (2020) Assessment of ethanol tolerance of *Kluyveromyces* marxianus CCT 7735 selected by adaptive laboratory evolution. Appl Microbiol Biotechnol 104:7483–7494 . https://doi.org/10.1007/s00253-020-10768-9

Silveira WB, Passos FJ V, Mantovani HC, Passos FML (2005) Ethanol production from cheese whey permeate by *Kluyveromyces marxianus* UFV-3: A flux analysis of oxido-reductive metabolism as a function of lactose concentration and oxygen levels. Enzyme Microb Technol 36:930–936 . https://doi.org/10.1016/j.enzmictec.2005.01.018

Sims REH, Mabee W, Saddler JN, Taylor M (2010) An overview of second generation biofuel technologies. Bioresour Technol 101:1570–1580 . https://doi.org/10.1016/j.biortech.2009.11.046

Šoštarić N, Arslan A, Carvalho B, Plech M, Voordeckers K, Verstrepen KJ, Van Noort V (2021) Integrated Multi-Omics Analysis of Mechanisms Underlying Yeast Ethanol Tolerance. J Proteome Res 20:3840–3852 . https://doi.org/10.1021/acs.jproteome.1c00139

Stanley D, Bandara A, Fraser S, Chambers PJ, Stanley GA (2010) The ethanol stress response and ethanol tolerance of *Saccharomyces cerevisiae*. J Appl Microbiol 109:13–24 . https://doi.org/10.1111/j.1365-2672.2009.04657.x

Stanley D, Fraser S, Chambers PJ, Rogers P, Stanley GA (2010) Generation and characterisation of stable ethanol-tolerant mutants of *Saccharomyces cerevisiae*. J

Ind Microbiol Biotechnol 37:139-149 https://doi.org/ 10.1007/s10295-009-0655-3

Steensels J, Snoek T, Meersman E, Nicolino MP, Voordeckers K, Verstrepen KJ (2014) Improving industrial yeast strains: Exploiting natural and artificial diversity. FEMS Microbiol Rev 38:947–995 . https://doi.org/10.1111/1574-6976.12073

Tavares B, Felipe M das G de A, dos Santos JC, Pereira FM, Gomes SD, Sene L (2019) An experimental and modeling approach for ethanol production by *Kluyveromyces marxianus* in stirred tank bioreactor using vacuum extraction as a strategy to overcome product inhibition. Renew Energy 131:261–267 . https://doi.org/10.1016/j.renene.2018.07.030

Tinôco D, Genier HLA, da Silveira WB (2021) Technology valuation of cellulosic ethanol production by *Kluyveromyces marxianus* CCT 7735 from sweet sorghum bagasse at elevated temperatures. Renew Energy 173:188–196 . https://doi.org/10.1016/j.renene.2021.03.132

Tokuyama K, Toya Y, Horinouchi T, Furusawa C, Matsuda F, Shimizu H (2018) Application of adaptive laboratory evolution to overcome a flux limitation in an *Escherichia coli* production strain. Biotechnol Bioeng 115:1542–1551 . https://doi.org/10.1002/bit.26568

Tomás-Pejó E, Oliva JM, González A, Ballesteros I, Ballesteros M (2009) Bioethanol production from wheat straw by the thermotolerant yeast *Kluyveromyces marxianus* CECT 10875 in a simultaneous saccharification and fermentation fed-batch process. Fuel 88:2142–2147 . https://doi.org/10.1016/j.fuel.2009.01.014

Trotter EW, Collinson EJ, Dawes IW, Grant CM (2006) Old yellow enzymes protect against acrolein toxicity in the yeast *Saccharomyces cerevisiae*. Appl Environ Microbiol 72:4885–4892 . https://doi.org/10.1128/AEM.00526-06

Vamvakas SS, Kapolos J (2020) Factors affecting yeast ethanol tolerance and fermentation efficiency. World J Microbiol Biotechnol 36:1–8 . https://doi.org/10.1007/s11274-020-02881-8

Vriesekoop F, Haass C, Pamment NB (2009) The role of acetaldehyde and glycerol in the adaptation to ethanol stress of *Saccharomyces cerevisiae* and other yeasts. FEMS Yeast Res 9:365–371. https://doi.org/10.1111/j.1567-1364.2009.00492.x

Wang D, Wu D, Yang X, Hong J (2018) Transcriptomic analysis of thermotolerant yeast: *Kluyveromyces marxianus* in multiple inhibitors tolerance. RSC Adv 8:14177–14192 . https://doi.org/10.1039/c8ra00335a

Wu WH, Hung WC, Lo KY, Chen YH, Wan HP, Cheng KC (2016) Bioethanol production from taro waste using thermo-tolerant yeast *Kluyveromyces marxianus* K21. Bioresour Technol 201:27–32 . https://doi.org/10.1016/j.biortech.2015.11.015

Yan J, Wei Z, Wang Q, He M, Li S, Irbis C (2015) Bioethanol production from sodium hydroxide/hydrogen peroxide-pretreated water hyacinth via simultaneous saccharification and fermentation with a newly isolated thermotolerant *Kluyveromyces marxianus* strain. Bioresour Technol 193:103–109 . https://doi.org/10.1016/j.biortech.2015.06.069

Yan T, Zhao Y (2020) Acetaldehyde induces phosphorylation of dynamin-related protein 1 and mitochondrial dysfunction via elevating intracellular ROS and Ca2+ levels. Redox Biol 28:101381 . https://doi.org/10.1016/j.redox.2019.101381

Yang KM, Lee NR, Woo JM, Choi W, Zimmermann M, Blank LM, Park JB (2012) Ethanol reduces mitochondrial membrane integrity and thereby impacts carbon metabolism of *Saccharomyces cerevisiae*. FEMS Yeast Res 12:675–684 . https://doi.org/10.1111/j.1567-1364.2012.00818.x

Yoshida M, Kato S, Fukuda S, Izawa S (2021) Acquired Resistance to Severe Ethanol Stress in *Saccharomyces cerevisiae* Protein Quality Control. Appl Environ Microbiol 87:1–17 . https://doi.org/10.1128/AEM.02353-20

Yu CY, Jiang BH, Duan KJ (2013) Production of bioethanol from carrot pomace using the thermotolerant yeast *Kluyveromyces marxianus*. Energies 6:1794–1801 . https://doi.org/10.3390/en6031794

Yuan WJ, Chang BL, Ren JG, Liu JP, Bai FW, Li YY (2012) Consolidated bioprocessing strategy for ethanol production from Jerusalem artichoke tubers by *Kluyveromyces marxianus* under high gravity conditions. J Appl Microbiol 112:38–44 . https://doi.org/10.1111/j.1365-2672.2011.05171.x

Zhao XQ, Bai FW (2009) Mechanisms of yeast stress tolerance and its manipulation for efficient fuel ethanol production. J Biotechnol 144:23–30 . https://doi.org/10.1016/j.jbiotec.2009.05.001

Zoppellari F, Bardi L (2013) Production of bioethanol from effluents of the dairy industry by *Kluyveromyces marxianus*. N Biotechnol 30:607–613 . https://doi.org/10.1016/j.nbt.2012.11.017

2.2 Multi-omics data and model integration reveal the main mechanisms associated with respiro-fermentative metabolism and ethanol stress responses in *Kluyveromyces marxianus*

Text adapted from the preprint version submitted to bioRxiv, available at: https://doi.org/10.1101/2024.06.06.597719.

Abstract

Kluyveromyces marxianus is a yeast capable of fermenting sugars into ethanol and growing at high temperatures (>37ºC). However, it is less tolerant to ethanol than *Saccharomyces cerevisiae*, which limits its application in second-generation ethanol production. Since the mechanisms of ethanol stress response are still poorly described, especially compared to *S. cerevisiae*, we used an integrative multi-omics approach, combining transcriptomics, co-expression networks, gene regulation, and genome-scale metabolic modelling to gain insights about these mechanisms. Through metabolic modelling, we predicted the occurrence of a respiro-fermentative metabolism and its onset as the dilution rate increased. From gene co-expression networks, we detected that the protein quality control system is a main mechanism involved in the ethanol stress response. Further, we identified key regulators in the ethanol stress response, such as *HAP3*, *MET4*, and *SNF2*, and assessed how disturbances in their gene expression affect cellular metabolism. We also found that amino acid metabolism, membrane lipid metabolism, and ergosterol exhibit increased metabolic flux under the explored conditions. These findings provide useful cues to develop and implement genetic and metabolic engineering strategies to enhance ethanol tolerance.

Keywords: Fermentation, metabolic modelling, enzyme allocation, gene regulation.

Highlights

- *Kluyveromyces marxianus* ferments sugars into ethanol and grows at high temperatures
- It's less ethanol-tolerant than *Saccharomyces cerevisiae*, limiting its industrial application

- Enzyme-constrained models predict the occurrence of respiro-fermentative metabolism
- Protein quality control is a key mechanism involved in the ethanol stress response
- Key regulators in ethanol stress response include *HAP3*, *MET4*, and *SNF2*

Introduction

Climate change has driven the development of new biofuel production processes to mitigate the damage caused by greenhouse gas emissions from burning fossil fuels [1–3]. Ethanol, the main biofuel marketed globally [4], is primarily produced from sugarcane and starch-based raw materials (first-generation ethanol) [3]. However, this process competes with food production. In this context, producing ethanol from alternative sources has been considered essential to avoid using arable land, water resources, and fertilisers, and to prevent competition with food production [5]. These alternative sources are generally industrial by-products, such as whey, which is rich in lactose, and lignocellulosic biomass, which is rich in glucose, xylose, mannose, and arabinose.

First-generation ethanol is produced by the yeast *Saccharomyces cerevisiae* [6] due to its remarkable ability to produce high ethanol titres from glucose and sucrose. At high sugar concentrations, enzymes of the central metabolism have their abundance altered, favouring enzymes that catalyse fermentative metabolism over respiratory metabolism [7]. This phenomenon, known as overflow metabolism (or the Crabtree effect in yeasts), allows *S. cerevisiae* to produce ethanol by fermentation even under aerobic conditions [8]. Despite its performance, *S. cerevisiae* is unable to assimilate sugars found in abundant feedstocks such as lactose and xylose [5]. Moreover, *S. cerevisiae* does not tolerate high temperatures, making difficult its application for second-generation ethanol production thorough processes of simultaneous saccharification and fermentation of lignocellulosic biomasses, which generally occur between 50 and 60ºC [9]. On the other hand, *Kluyveromyces marxianus* is a thermotolerant yeast capable of fermenting sugars at temperatures ranging from 42 to 48ºC [10]. Additionally, it can use lactose and xylose as carbon and energy sources. Although it can produce ethanol from sugars such as lactose and glucose, it is not considered a Crabtree-positive yeast [11]. In addition, *K. marxianus* is less tolerant to high ethanol

concentrations, which is a drawback for its application in ethanol production [12,13]. Thus, efforts have been made to uncover the mechanisms of ethanol stress response in this yeast and to obtain strains with improved tolerance [14].

Studies on ethanol stress in yeasts have primarily focused on *S. cerevisiae*. There are still few studies on non-*Saccharomyces* yeasts. In *K. marxianus*, physiological responses to stress conditions are largely inferred from results with its sister species, *K. lactis*, and with *S. cerevisiae* [14]. However, there are fundamental differences between the species, highlighting the need for a better understanding of stress response mechanisms in *K. marxianus*. Specifically, during ethanol stress, *S. cerevisiae* induces an increase in ergosterol and unsaturated fatty acids in the plasma membrane, while in *K. marxianus*, this response relies on the strain and cultivation condition [15]. Regarding gene expression, different sets of genes are induced and repressed between the two species. Differences in metabolic profiles between *K. marxianus* and *S. cerevisiae* are also notable [16]. Additionally, there are few large-scale studies at the systems level that integrate both molecular and metabolic knowledge.

Metabolism can be studied computationally through flux balance analysis (FBA) where genome-scale metabolic models (GEMs) are used to predict metabolic flux [17]. GEMs include the stoichiometry of each reaction in the network and information about the genes responsible for the enzymes that catalyse each reaction, allowing a direct association between genotype and phenotype. These models are useful for phenotype simulations and can be improved with additional data, especially catalytic efficiency and enzyme concentration [18,19], and gene expression [20]. GEMs integrated with enzyme constraints (ecGEMs) enable the study of proteome resource allocation by considering enzyme concentration and allow the study of more complex phenotypes, such as metabolic changes that occur at high growth rates and carbon source concentrations. Specifically, ecGEMs allow studying and predicting the Crabtree effect, which occurs in a manner dependent on the allocation of enzymatic resources. In *K. marxianus*, which exhibits respiro-fermentative metabolism [10], there are currently no modelling studies on this respiro-fermentative metabolism. Despite the utility of GEMs and ecGEMs, metabolism is highly dependent on gene expression and regulation, which can be studied through gene co-expression networks (GCNs) and gene regulatory networks (GRNs). These networks can be inferred from transcriptomics data, associating expression patterns of individual genes that occur jointly (in the case of GCNs) [21], or

associating the expression of transcription factor (TF) genes with possible target genes through unsupervised or reference-guided clustering (in the case of GRNs) [22].

To address this knowledge gap in *K. marxianus*, here we focus on reconstructing and analysing the enzyme-constrained metabolic network of *K. marxianus*, and characterising GCNs and GRNs using available transcriptomics data. Further, we integrate these different modelling approaches to investigate gene co-expression, gene regulation, metabolism regulation, and resource allocation in response to ethanol stress. Finally, we were able to predict the respiro-fermentative metabolism *K. marxianus* and establish a basis for proposing metabolic engineering strategies to improve tolerance to high ethanol concentrations in *K. marxianus*.

Material and methods

Integrating enzyme constraints in the K. marxianus GEM

To enhance the predictive capacity of the GEM for *K. marxianus*, we integrated catalytic efficiency values (k_{cat}) and enzyme concentrations into the consensus model iSM996 [23]. For this, we used the GECKO Toolbox 3 [24]. Information about the enzymes in the model, such as molecular mass, EC number, and amino acid sequence, was retrieved from UniProt [25] and KEGG [26]. The k_{cat} values were obtained in two ways: using values included in the BRENDA database [27] and predicted by the DLKcat tool [28]. For each k_{cat} value of an enzyme obtained from BRENDA or DLKcat, the highest available value was used. To make predictions using DLKcat, the SMILES code of each metabolite was obtained from the PubChem database [29]. Lastly, we allowed flexibilization of the k_{cat} values for the model to achieve the maximum growth rate of 0.56 h^{-1} using glucose as carbon source. To evaluate the predictive capabilities of the model, we simulated batch cultures using different carbon sources (glucose, fructose, sucrose, galactose, lactose, and xylose) and compared the predictions to experimental data [30–36], with the following optimization set-up:

$$max\ v_{bio} \tag{P1}$$

subject to

$$\boldsymbol{N} \cdot \boldsymbol{v} = 0 \tag{1}$$

$$v_j^{lb} \leq v_j \leq v_j^{ub} \quad (2)$$

$$v_j \leq k_{cat}^{i,j} \cdot E_i \quad (3)$$

where v_{bio} is the flux through the biomass pseudoreaction, $\mathbf{N}$ is the stoichiometric matrix, $\mathbf{v}$ is the flux distribution vector, v_j^{lb} is the flux lower bound of reaction j, v_j is the flux through reaction j, v_j^{ub} is the flux upper bound of reaction j, $k_{cat}^{i,j}$ is the k_{cat} of enzyme i catalyzing reaction j, and E is the concentration of enzyme i.

Next, we simulated glucose-limited chemostats with increasing dilution and substrate uptake rates to predict metabolic shifts. We constrained v_{bio} with respect to increasing values of 0 to 0.5, while minimizing the flux through the carbon source exchange reaction:

$$min\, v_{carbon} \quad (P2)$$

subject to

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

$$v_j^{lb} \leq v_j \leq v_j^{ub} \quad (2)$$

$$v_j \leq k_{cat}^{i,j} \cdot E_i \quad (3)$$

$$v_{bio} = \mu \quad (4)$$

where v_{carbon} is the flux through the carbon source exchange reaction, and μ is the growth rate. We named the finalized model as eciSM996. All model refinements were performed using the COBRA Toolbox 3 [37] and/or the RAVEN Toolbox 2 [38] in MATLAB (The MathWorks Inc., Natick, Massachusetts).

Obtaining and processing RNAseq data

To investigate the effects of ethanol stress in *K. marxianus*, we used transcriptomic data from Diniz et al. [39] and Mo et al. [40], which were obtained from cultures of *K. marxianus* exposed to ethanol. The data obtained from Diniz et al. [39] were unprocessed FASTQ reads, while Mo et al. [40] provided the gene count matrix. To process the data from Diniz et al. [39], we aligned the reads to the *K. marxianus* reference genome (strain DMKU3-1042) using Bowtie2 [41]. Using these alignments, we generated the count matrix using featureCounts [42]. Lastly, we normalised the counts with

DESeq2 [43], using the geometric mean for each gene across all samples, for both the counts generated from Diniz et al. [39] and the matrix obtained from Mo et al. [40].

Inferring gene co-expression and regulation networks

To identify key genes in the ethanol stress response, we first inferred gene co-expression networks (GCNs). For this, we used the BioNERO package [44], available in R/Bioconductor [45]. We used the normalised count matrices as input data and filtered out genes with no expression value and outlier genes. We also removed genes that function as confounding variables. The inferred GCNs were of the hybrid signed type, so only positive associations in the adjacency matrix are considered:

$$a_{ij} = \left[cor\ (x_i, x_j)\right]^{\beta}, for\ \left(x_i, x_j\right) > 0 \quad (5)$$

$$a_{ij} = 0, for\ \left(x_i, x_j\right) \leq 0 \quad (6)$$

where a is the i-th, j-th element of the adjacency matrix, x is the expression value of the gene, and β is soft power threshold. Using the identified modules, we annotated them using Gene Ontology (GO) [46] and InterPro functional domains [47]. We selected the inferred modules with the highest correlation with the analysed phenotypes to identify the key genes.

We reconstructed the gene regulatory network (GRN) of *K. marxianus* by concatenating two GRNs. First, we inferred a GRN using BioNERO, with the normalised count matrices as input, identifying specific regulator-target pairs for each input matrix. Then, we generated a second GRN from the *S. cerevisiae* GRN available in the Yeastract database [48]. For each regulator and its respective target genes, we identified the corresponding orthologous genes in *K. marxianus*, substituting the existing gene families in both species and using bidirectional BLAST for genes without a direct association.

Integration of GRNs and metabolic models

To integrate the inferred GRNs with metabolism, we used the Probabilistic Regulation of Metabolism (PROM) algorithm [49]. PROM calculates the probability of a gene being regulated in the absence of its regulator. The relationship between regulator and target is binarized to represent "on" or "off" states for a given target. This probability is represented as:

$$P(A = 1|B = 0) = \frac{N(A = 1|B = 0)}{N(B = 0)} \quad (7)$$

where P is the probability, A is the target gene, B is the regulator, and N is the number of occurrences. We used PROM to constrain the conventional iSM996 model by solving the following problem:

$$\max Z = \sum_j c_j v_j + \sum_j (\kappa_j \alpha_h + \kappa_j \beta_j) \quad (P3)$$

subject to

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

$$P \cdot v_j^{lb} - \alpha_j \leq v_j \leq P \cdot v_j^{ub} - \beta_j \quad (8)$$

$$\alpha_j, \beta_j \geq 0 \quad (9)$$

where c is the objective coefficient for reaction j, α and β are adjustments of the lower bound v_j^{lb} and upper bound v_j^{ub}, and P is the probability calculated in Equation 3. We performed all simulations using the COBRA Toolbox 3 [37] and/or the RAVEN Toolbox 2 [38] in MATLAB (The MathWorks Inc., Natick, Massachusetts).

Integrating gene expression in the K. marxianus GEM

To investigate how variations in gene expression affect metabolism, we integrated gene expression data from the count matrices with the metabolic model iSM996 using RIPTiDe [50]. This tool minimises fluxes weighted by the expression of genes controlling the reactions:

$$\min Z = \sum v_{irrev,j} \cdot rw_{pruning} \quad (P4)$$

subject to

$$v_{bio} = v_{bio,max} \cdot f \quad (10)$$

$$\mathbf{N_{irrev}} \cdot \mathbf{v_{irrev}} = 0 \quad (11)$$

$$0 \leq \mathbf{v_{irrev}} \leq \mathbf{v_{max}} \quad (12)$$

where $v_{irrev,j}$ is the metabolic flux through an irreversible reaction j, $rw_{pruning}$ is the weight determined by gene expression, f is the correction factor, $\mathbf{N_{irrev}}$ is the stoichiometric matrix in irreversible format, and $\mathbf{v_{irrev}}$ is the flux vector for irreversible reactions. By obtaining condition-specific models, we maximised biomass production through the reduced network:

$$\max v_{bio} \quad \text{(P5)}$$

using constraints from Equations 1 and 2.

Next, we applied the predicted flux distribution as constraints in the eciSM996 model to predict resource allocation under these conditions. With this model, we solved problem P5, adding constraint Equation 3 and the following constraint:

$$v_j = v_{j,RIPTiDe} \; \forall \; v_j \in \; v_{j,RIPTiDe} \quad (13)$$

where $v_{j,RIPTiDe}$ is the previously predicted flux through reaction j.

Results and discussion

The ecGEM eciSM966 predicts respiro-fermentative metabolism

Conventional metabolic models are useful for predicting various phenotypes and physiological conditions, but they cannot simulate metabolic changes such as overflow metabolism or diauxic growth without additional constraints [19,51]. These phenotypes can be predicted by integrating enzymatic constraints into these models, such as k_{cat} values and enzyme concentrations. In this regard, we used GECKO3 to reconstruct the ecGEM of *K. marxianus* from the conventional GEM iSM966, resulting in eciSM966. This new model has 10,597 reactions, 2,523 metabolites, and 997 genes. Out of the total reactions, 9,293 of them have integrated k_{cat}values, distributed among 991 enzymes.

Using this model, we simulated batch cultures using different carbon sources. We compared the predicted growth rates to experimental growth rates and those predicted by a conventional GEM (Figure 1).

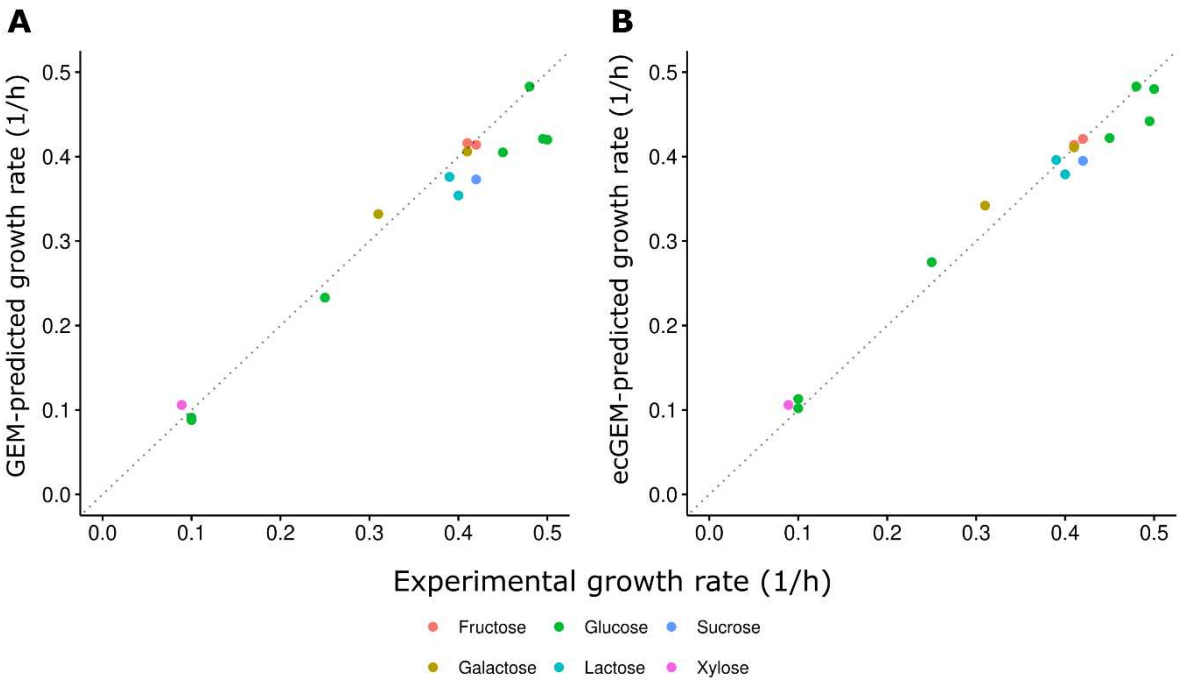

Figure 1. Growth rate comparison between experimental values and predicted values. (A) GEM predictions. (B) ecGEM predictions.

We observed that the eciSM966 model can predict growth rates closer to experimental values than the conventional GEM. Then, we simulated chemostats at different dilution rates to assess metabolic changes associated with high growth rates. We observed that glucose and oxygen uptake, and CO_2 production, linearly increase along dilution rates up to a rate of 0.35 h^{-1} (Figure 2).

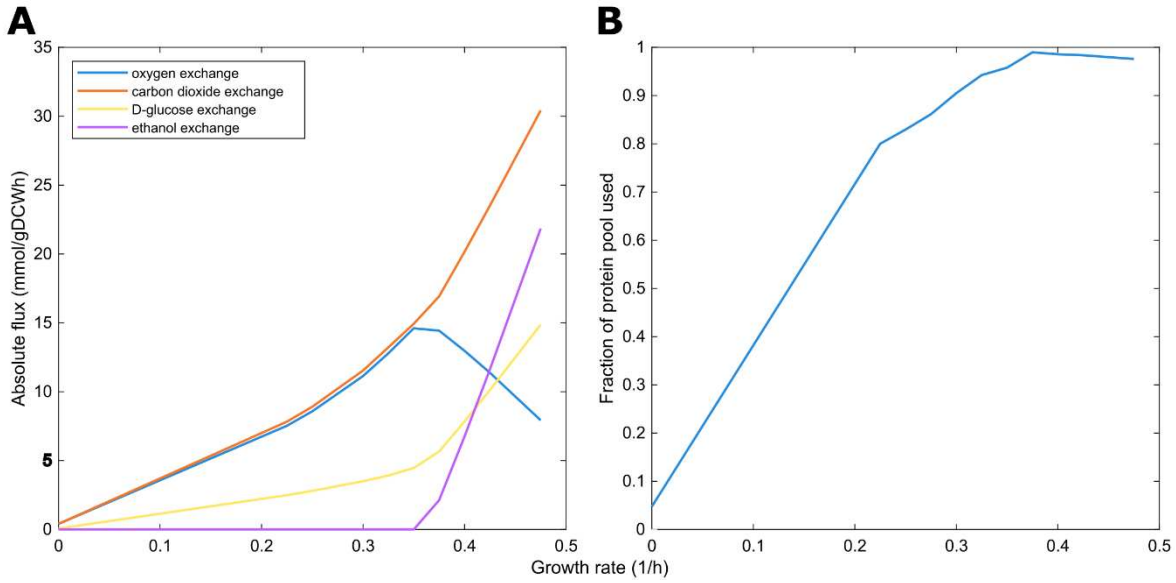

Figure 2. Simulation of chemostats at increasing dilution rates. (A) Absolute flux for oxygen and glucose uptake, and CO2 and ethanol production. (B) Fraction of protein resource use as a function of dilution rate.

From this rate onwards, there is a spike in ethanol production and a decrease in oxygen uptake. However, oxygen continues to be taken up, unlike what occurs in the ecYeastGEM model of *S. cerevisiae* [19]. Additionally, the fraction of total available proteins used also increased up to a dilution rate of 0.35 h^{-1}, with a slight decrease after this rate, indicating less enzyme allocation at higher dilution rates. These results correlate with those reported on overflow metabolism, where different sets of enzymes are activated and repressed in this metabolic shift, employing enzymes with higher catalytic efficiency to deal with fermentative metabolism. Nevertheless, the observed oxygen uptake at high dilution rates indicates that *K. marxianus*, contrary to *S. cerevisiae* (Crabtree-positive yeast), exhibits a respiro-fermentative type of overflow metabolism, where respiratory and fermentative pathways occur simultaneously at high growth rates. This is agreement with high-sugar batch cultivations of *K. marxianus* conducted in Erlenmeyer flasks, where the cultivations under hypoxia displayed higher lactose consumption and ethanol production than anoxic cultivations, highlighting that lower but still present oxygen content favours the ethanol production from different feedstocks [13,52–55].

Gene co-expression networks reveals that the protein quality control system displays a key role in response to ethanol stress

To investigate the mechanisms affected in the response to ethanol stress in *K. marxianus*, we inferred gene co-expression networks to identify gene modules related to physiological changes. First, we collected transcriptomic data from the literature, using data generated from yeast cultivated under unstressed conditions and exposed to ethanol (stress condition), from two different studies [39,40]. The data from each study were independently used for inferring the GCNs.

In the study conducted by Diniz et al. [39], triplicate samples were collected at three different time intervals: zero hours after 6% (v/v) ethanol exposure (0h, control), one hour after exposure (1h), and four hours after exposure (4h). With the normalized gene count matrix, we first determined the β power value that best satisfied a scale-free topology without drastically reducing average connectivity (Figure S1). The optimal network presented 33 modules (Figure 3). We then performed a Gene Ontology (GO) enrichment analysis for modules positively correlated with the groups. For the 0h condition, the most correlated module contained genes associated with ribosome biogenesis activity (Figure S2). For the 1h condition, the modules featured genes related to glycosylphosphatidylinositol (GPI) anchoring to proteins, nucleolus activity, and ribosomal RNA (rRNA) processing (Figure S3). For the 4h condition, the modules contained genes related to ribosome structure and protein biosynthesis activity (Figure S4).

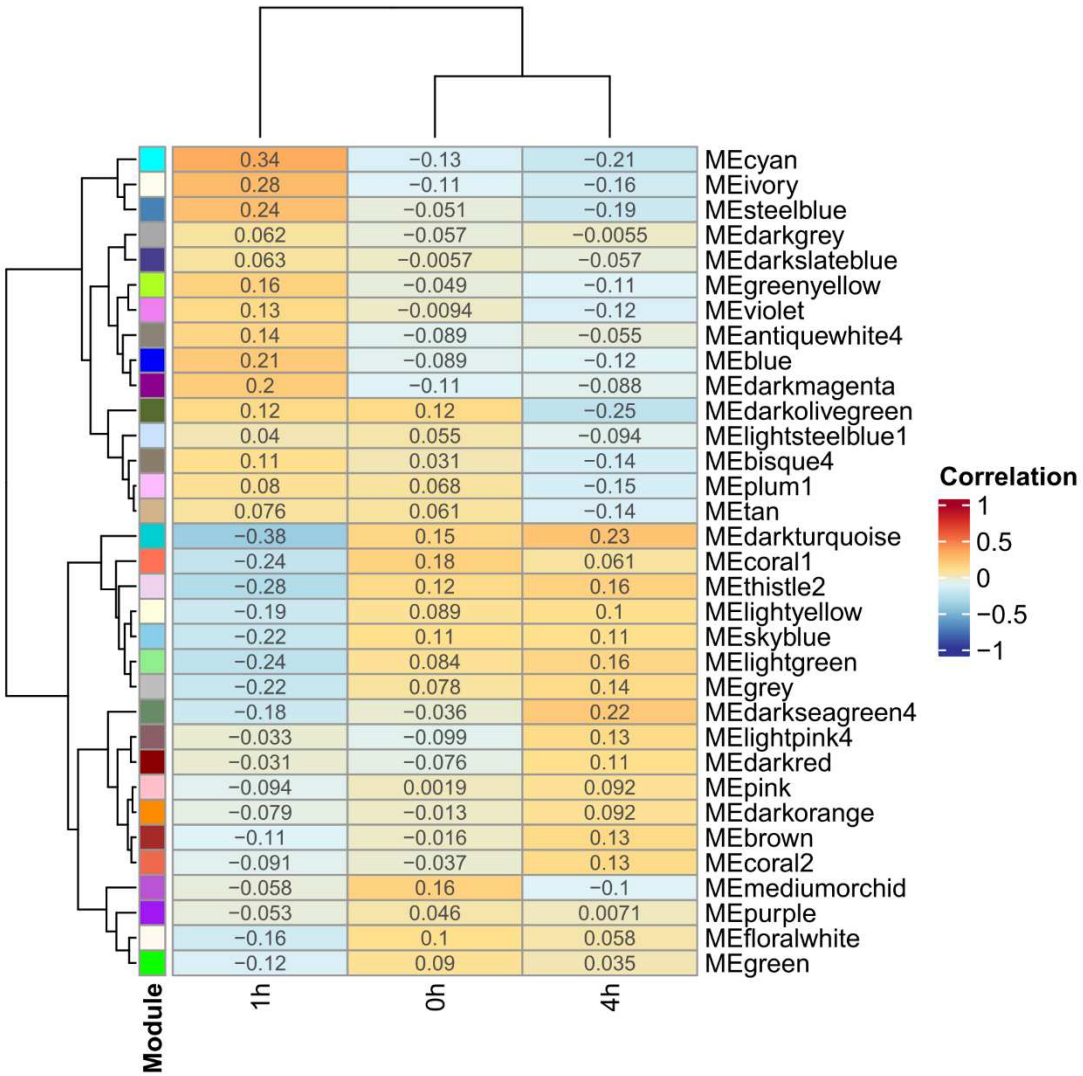

Figure 3. Modules detected and their correlation with the phenotype analysed for the data from Diniz et al. [39].

In the study of Mo et al. [40], laboratory adaptive evolution (ALE) was performed in *K. marxianus* exposed to 6% ethanol (v/v) for 450 generations in order to select evolved strains with improved tolerance to the ethanol stress. For evaluation of the evolved strain, triplicate samples were collected from the following cultures: 0% ethanol (0%-KM, control), 4% ethanol (4%-KM), and 6% ethanol (6%-KM), using the wild-type strain; and 0% ethanol (0%-100d), 4% ethanol (4%-100d), and 6% ethanol (6%-100d) using the evolved strain with higher ethanol tolerance. For inference of the

GCNs, we used the count matrix to determine the β power value (Figure S5). The optimal network presented 27 modules (Figure 4).

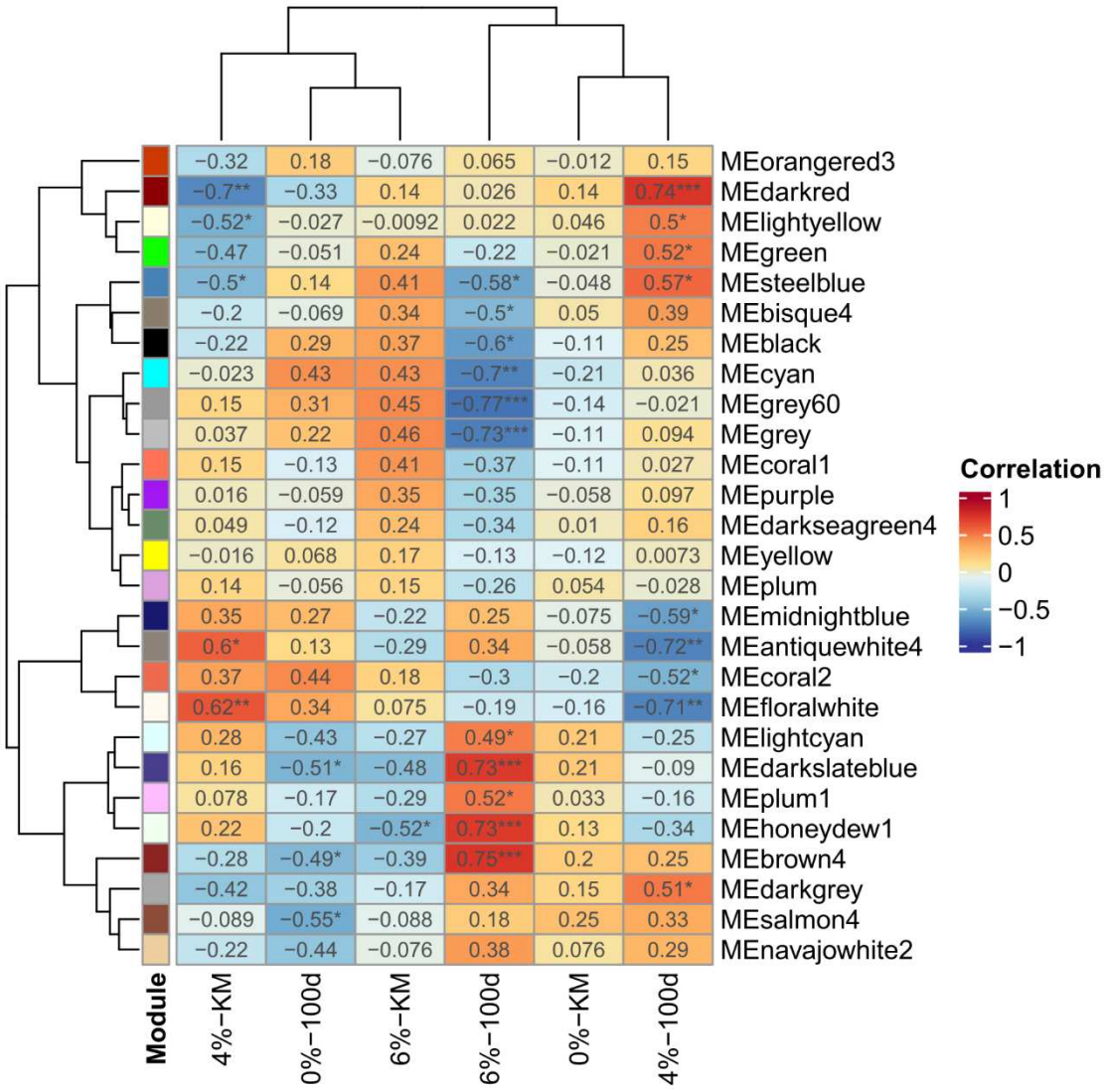

Figure 4. Modules detected and their correlation with the phenotype analysed for the data from Mo et al. [40]

We analysed GO enrichment for modules positively correlated with the groups. For the 0%-KM condition, there were no modules with enriched functions. For 4%-KM, the modules featured genes involved in the citric acid cycle (CAC) and inner membrane of the ribosome (Figure S6). In the 6%-KM condition, the modules contained genes involved in protein transport, post-transcriptional and post-translational modifications,

and ubiquitination (Figure S7). The 0%-100d condition also had a module containing genes involved in these same activities (Figure S7A). In the 4%-100d condition, the modules contained genes involved in proteasome activities and protein ubiquitination, microtubules, and kinase regulators (Figure S8). Finally, in the 6%-100d condition, the modules contained genes involved in ribosomal structure and function, protein biosynthesis, and ribosomes (Figure S9). Together, these results indicate that there is a remodelling of the gene expression machinery during ethanol stress, as the modules positively correlated with this condition were mostly enriched with genes related to transcription, translation, and regulation activities. Additionally, in the mutant strains of Mo et al. [40], we observed the presence of genes related to the proteasome, suggesting that protein degradation and recycling are affected during ethanol stress. Since ethanol causes protein denaturation, this could explain why genes related to the gene expression machinery were enriched by certain GO terms, indicating that the protein quality control system is a pivotal mechanism in response to the ethanol stress in *K. marxianus*.

Metabolic impact of transcription factor knockouts

Using the inferred GRNs, we integrated the networks with the iSM996 GEM using PROM to predict how the knockout of each regulator individually impacts metabolism, investigating changes in specific growth rates and flux distributions from the different samples of Diniz et al. [39] and Mo et al. [40].

For the Diniz et al. [39] samples, we observed a total of 13 regulators whose knockout led to a decrease in specific growth rate (Table S1). Among these, notable ones include: the transcriptional activator *HAP3* in the 4h condition, which regulates genes involved in the electron transport chain in the mitochondria; and the transcriptional activator *MET4*, involved in regulating genes for the biosynthesis of sulphur-containing amino acids (methionine, cysteine, and cystine). For the Mo et al. [40] samples, the deletion of all transcription factors resulted in an impact on the specific growth rate (Table S2); notably, the knockout of the regulator *SNF2* led to the lowest specific growth rate among the regulators. This regulator is part of a family of helicase proteins that play a role in chromatin remodelling and is involved in molecular responses to amino acid depletion, response to double-strand DNA breaks, and responses to glucose depletion, all triggered during ethanol stress [14].

Condition-specific models reveal flux redistributions and enzyme activities

To investigate how changes in gene expression affect metabolism, we integrated transcriptomic data with the iSM966 model using RIPTiDe, which restricts the flux of reactions controlled by genes present in the count matrix. We generated condition-specific models for each cultivation condition, from both the Diniz et al. [39] and Mo et al. [40] data. With these condition-specific models, we predicted the corresponding flux distribution, which we then used to constrain the model with enzymatic restriction, eciSM966. With a new round of predictions, we were able to predict the specific growth rate and obtain the flux distribution that approximates to the organism's physiological reality given the measured gene expression.

The condition-specific models obtained from the Diniz et al. [39] and Mo et al. [40] samples reveal similar phenotypes when compared with the eciSM966 model. Focusing on samples with higher ethanol exposure (6h, 6-KM, 6-100d), we observed that in all three situations, the specific growth rate is low ($\mu \leq 0.07$). Regarding the flux distribution, metabolic flux was predicted for several reactions related to ethanol metabolism, involving enzymes such as acetaldehyde:NAD+ oxidoreductase and ethanol:NAD+ oxidoreductase. In the 6h condition specifically, we predicted flux for branched-chain amino acid and amino acid production, while in the 6-KM and 6-100d conditions, flux was predicted for ergosterol and membrane lipid biosynthesis. This result is consistent with the results previously described experimentally in the works of Diniz et al. [39] and Mo et al. [40]. Further, the results from the 6-KM and 6-100d conditions also correlate to the ALE experiments of Silveira et al. [56], which have also detected increases in ergosterol and membrane lipid contents in the evolved strains. These evolved strains have also shown an increase in valine (a branched-chain amino acid) and metabolites involved in central pathways, such as isocitric acid, citric acid and cis-aconitic acid, for which the 6h condition-specific model could successfully predict.

Conclusion

Here we predicted the occurrence of the respiro-fermentative metabolism *in K. marxianus* by using enzyme-constrained metabolic models. Moreover, we obtained new insights about the genes, enzymes, and metabolites involved in the response to ethanol

stress in this yeast, integrating transcriptomic data with co-expression network modelling approaches, gene regulatory networks, and metabolic networks. By obtaining results at the resolution of metabolic flux, we identified how metabolism is remodelled to ensure cell survival and the impacts on cell growth. Thus, these results indicate target genes for the implementation of strategies to increase ethanol tolerance, paving the way for further studies of metabolic engineering.

Acknowledgments

This study was financed in part by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior – Brasil (CAPES) – Finance Code 001 and Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq) – Process 312390/2020-3. We thank Eduardo Almeida for his discussions and critical comments on this study.

Data availability

The code and data used are publicly available in the GitHub repository: https://github.com/LabFisUFV/KmarxianusEthanol. A static version of the repository is available at Zenodo: https://zenodo.org/doi/10.5281/zenodo.11501521.

References

[1] S.I. Mussatto, G. Dragone, P.M.R. Guimarães, J.P.A. Silva, L.M. Carneiro, I.C. Roberto, A. Vicente, L. Domingues, J.A. Teixeira, Technological trends, global market, and challenges of bio-ethanol production, Biotechnology Advances 28 (2010) 817–830. https://doi.org/10.1016/j.biotechadv.2010.07.001.

[2] M. Unglert, D. Bockey, C. Bofinger, B. Buchholz, G. Fisch, R. Luther, M. Müller, K. Schaper, J. Schmitt, O. Schröder, U. Schümann, H. Tschöke, E. Remmele, R. Wicht, M. Winkler, J. Krahl, Action areas and the need for research in biofuels, Fuel 268 (2020) 117227. https://doi.org/10.1016/j.fuel.2020.117227.

[3] M. Valdivia, J.L. Galan, J. Laffarga, J.L. Ramos, Biofuels 2020: Biorefineries based on lignocellulosic materials, Microbial Biotechnology 9 (2016) 585–594. https://doi.org/10.1111/1751-7915.12387.

[4] RFA, 2020 RFA's Ethanol Industry Outlook, 2020.

[5] K. Robak, M. Balcerek, Review of second generation bioethanol production from residual biomass, Food Technology and Biotechnology 56 (2018) 174–187. https://doi.org/10.17113/ftb.56.02.18.5428.

[6] I.E.A. IEA, G. Cooper, J. McCaherty, E. Huschitt, R. Schwarck, C. Wilson, 2021 Ethanol Industry Outlook, Renewable Fuels Association (2021) 1–40.

[7] S. Moreno-Paz, J. Schmitz, V.A.P. Martins dos Santos, M. Suarez-Diez, Enzyme-constrained models predict the dynamics of Saccharomyces cerevisiae growth in continuous, batch and fed-batch bioreactors, Microbial Biotechnology 15 (2022) 1434–1445. https://doi.org/10.1111/1751-7915.13995.

[8] E. de Alteriis, F. Cartenì, P. Parascandola, J. Serpa, S. Mazzoleni, Revisiting the Crabtree/Warburg effect in a dynamic perspective: a fitness advantage against sugar-induced cell death, Cell Cycle 17 (2018) 688–701. https://doi.org/10.1080/15384101.2018.1442622.

[9] Y.J. Liu, B. Li, Y. Feng, Q. Cui, Consolidated bio-saccharification: Leading ligno-cellulose bioconversion into the real world, Biotechnology Advances 40 (2020) 107535. https://doi.org/10.1016/j.biotechadv.2020.107535.

[10] M.M. Lane, J.P. Morrissey, Kluyveromyces marxianus: A yeast emerging from its sister's shadow, Fungal Biology Reviews 24 (2010) 17–26. https://doi.org/10.1016/j.fbr.2010.01.001.

[11] M. Arellano-Plaza, R. Noriega-Cisneros, M. Clemente-Guerrero, J.C. González-Hernández, P.D. Robles-Herrera, S. Manzo-Ávalos, A. Saavedra-Molina, A.

Gschaedler-Mathis, Fermentative capacity of Kluyveromyces marxianus and Saccharomyces cerevisiae after oxidative stress, Journal of the Institute of Brewing 123 (2017) 519–526. https://doi.org/10.1002/jib.451.

[12] D.A. Costa, C.J.A. de Souza, P.A. Costa, M.Q.R.B. Rodrigues, A.F. dos Santos, M.R. Lopes, H.L.A. Genier, W.B. Silveira, L.G. Fietto, Physiological characterization of thermotolerant yeast for cellulosic ethanol production., Applied Microbiology and Biotechnology 98 (2014) 3829–40. https://doi.org/10.1007/s00253-014-5580-3.

[13] W.B. Silveira, F.J.V. Passos, H.C. Mantovani, F.M.L. Passos, Ethanol production from cheese whey permeate by Kluyveromyces marxianus UFV-3: A flux analysis of oxido-reductive metabolism as a function of lactose concentration and oxygen levels, Enzyme and Microbial Technology 36 (2005) 930–936. https://doi.org/10.1016/j.enzmictec.2005.01.018.

[14] M.A. de Moura Ferreira, F.A. da Silveira, W.B. da Silveira, Ethanol stress responses in Kluyveromyces marxianus: current knowledge and perspectives, Appl Microbiol Biotechnol 106 (2022) 1341–1353. https://doi.org/10.1007/s00253-022-11799-0.

[15] X. Fu, P. Li, L. Zhang, S. Li, Understanding the stress responses of Kluyveromyces marxianus after an arrest during high-temperature ethanol fermentation based on integration of RNA-Seq and metabolite data, Applied Microbiology and Biotechnology 103 (2019) 2715–2729. https://doi.org/10.1007/s00253-019-09637-x.

[16] F. Vriesekoop, C. Haass, N.B. Pamment, The role of acetaldehyde and glycerol in the adaptation to ethanol stress of Saccharomyces cerevisiae and other yeasts,

FEMS Yeast Research 9 (2009) 365–371. https://doi.org/10.1111/j.1567-1364.2009.00492.x.

[17] J.D. Orth, I. Thiele, B.Ø. Palsson, What is flux balance analysis?, Nature Biotechnology 28 (2010) 245–248. https://doi.org/10.1038/nbt.1614.

[18] R. Adadi, B. Volkmer, R. Milo, M. Heinemann, T. Shlomi, Prediction of Microbial Growth Rate versus Biomass Yield by a Metabolic Network with Kinetic Parameters, PLOS Computational Biology 8 (2012) e1002575. https://doi.org/10.1371/JOURNAL.PCBI.1002575.

[19] B.J. Sánchez, C. Zhang, A. Nilsson, P.-J. Lahtvee, E.J. Kerkhoven, J. Nielsen, Improving the phenotype predictions of a yeast genome-scale metabolic model by incorporating enzymatic constraints, Molecular Systems Biology 13 (2017) 935. https://doi.org/10.15252/MSB.20167411.

[20] W. Guo, X. Feng, OM-FBA: Integrate Transcriptomics Data with Flux Balance Analysis to Decipher the Cell Metabolism, PLOS ONE 11 (2016) e0154188. https://doi.org/10.1371/journal.pone.0154188.

[21] C. Ruprecht, N. Vaid, S. Proost, S. Persson, M. Mutwil, Beyond Genomics: Studying Evolution with Gene Coexpression Networks, Trends in Plant Science 22 (2017) 298–307. https://doi.org/10.1016/j.tplants.2016.12.011.

[22] M. Zhao, W. He, J. Tang, Q. Zou, F. Guo, A comprehensive overview and critical evaluation of gene regulatory network inference technologies, Briefings in Bioinformatics 00 (2021) 1–15. https://doi.org/10.1093/bib/bbab009.

[23] S. Marcišauskas, B. Ji, J. Nielsen, Reconstruction and analysis of a Kluyveromyces marxianus genome-scale metabolic model, BMC Bioinformatics 20 (2019) 1–9. https://doi.org/10.1186/s12859-019-3134-5.

[24] Y. Chen, J. Gustafsson, A. Tafur Rangel, M. Anton, I. Domenzain, C. Kittikunapong, F. Li, L. Yuan, J. Nielsen, E.J. Kerkhoven, Reconstruction, simulation and analysis of enzyme-constrained metabolic models using GECKO Toolbox 3.0, Nat Protoc 19 (2024) 629–667. https://doi.org/10.1038/s41596-023-00931-7.

[25] T.U. Consortium, UniProt: a worldwide hub of protein knowledge, Nucleic Acids Research 47 (2018) D506–D515. https://doi.org/10.1093/nar/gky1049.

[26] M. Kanehisa, S. Goto, KEGG: Kyoto Encyclopedia of Genes and Genomes, Nucleic Acids Research 28 (2000) 27–30. https://doi.org/10.1093/nar/28.1.27.

[27] A. Chang, L. Jeske, S. Ulbrich, J. Hofmann, J. Koblitz, I. Schomburg, M. Neumann-Schaal, D. Jahn, D. Schomburg, BRENDA, the ELIXIR core data resource in 2021: new developments and updates, Nucleic Acids Research 49 (2021) D498–D508. https://doi.org/10.1093/NAR/GKAA1025.

[28] F. Li, L. Yuan, H. Lu, G. Li, Y. Chen, M.K.M. Engqvist, E.J. Kerkhoven, J. Nielsen, Deep learning-based kcat prediction enables improved enzyme-constrained model reconstruction, Nature Catalysis (2022) 1–11. https://doi.org/10.1038/s41929-022-00798-z.

[29] S. Kim, J. Chen, T. Cheng, A. Gindulyte, J. He, S. He, Q. Li, B.A. Shoemaker, P.A. Thiessen, B. Yu, L. Zaslavsky, J. Zhang, E.E. Bolton, PubChem 2023 update, Nucleic Acids Research 51 (2023) D1373–D1380. https://doi.org/10.1093/nar/gkac956.

[30] G.G. Fonseca, N.M.B. de Carvalho, A.K. Gombert, Growth of the yeast Kluyveromyces marxianus CBS 6556 on different sugar combinations as sole carbon and energy source, Appl Microbiol Biotechnol 97 (2013) 5055–5067. https://doi.org/10.1007/s00253-013-4748-6.

[31] L.H. Bellaver, N.M.B. de Carvalho, J. Abrahão-Neto, A.K. Gombert, Ethanol formation and enzyme activities around glucose-6-phosphate in Kluyveromyces marxianus CBS 6556 exposed to glucose or lactose excess, FEMS Yeast Res 4 (2004) 691–698. https://doi.org/10.1016/j.femsyr.2004.01.004.

[32] G.G. Fonseca, A.K. Gombert, E. Heinzle, C. Wittmann, Physiology of the yeast Kluyveromyces marxianus during batch and chemostat cultures with glucose as the sole carbon source, FEMS Yeast Research 7 (2007) 422–435. https://doi.org/10.1111/j.1567-1364.2006.00192.x.

[33] E. Postma, P.J. Van den Broek, Continuous-culture study of the regulation of glucose and fructose transport in Kluyveromyces marxianus CBS 6556, Journal of Bacteriology 172 (1990) 2871–2876. https://doi.org/10.1128/jb.172.6.2871-2876.1990.

[34] A. Pentjuss, E. Stalidzans, J. Liepins, A. Kokina, J. Martynova, P. Zikmanis, I. Mozga, R. Scherbaka, H. Hartman, M.G. Poolman, D.A. Fell, A. Vigants, Model-based biotechnological potential analysis of Kluyveromyces marxianus central metabolism, Journal of Industrial Microbiology and Biotechnology 44 (2017) 1177–1190. https://doi.org/10.1007/s10295-017-1946-8.

[35] L.G.S. Longhi, D.J. Luvizetto, L.S. Ferreira, R. Rech, M.A.Z. Ayub, A.R. Secchi, A growth kinetic model of Kluyveromyces marxianus cultures on cheese whey as substrate, J IND MICROBIOL BIOTECHNOL 31 (2004) 35–40. https://doi.org/10.1007/s10295-004-0110-4.

[36] M. Hensing, H. Vrouwenvelder, C. Hellinga, R. Baartmans, H. van Dijken, Production of extracellular inulinase in high-cell-density fed-batch cultures of Kluyveromyces marxianus, Appl Microbiol Biotechnol 42 (1994) 516–521. https://doi.org/10.1007/BF00173914.

[37] L. Heirendt, S. Arreckx, T. Pfau, S.N. Mendoza, A. Richelle, A. Heinken, H.S. Haraldsdóttir, J. Wachowiak, S.M. Keating, V. Vlasov, S. Magnusdóttir, C.Y. Ng, G. Preciat, A. Žagare, S.H.J. Chan, M.K. Aurich, C.M. Clancy, J. Modamio, J.T. Sauls, A. Noronha, A. Bordbar, B. Cousins, D.C. El Assal, L.V. Valcarcel, I. Apaolaza, S. Ghaderi, M. Ahookhosh, M. Ben Guebila, A. Kostromins, N. Sompairac, H.M. Le, D. Ma, Y. Sun, L. Wang, J.T. Yurkovich, M.A.P. Oliveira, P.T. Vuong, L.P. El Assal, I. Kuperstein, A. Zinovyev, H.S. Hinton, W.A. Bryant, F.J. Aragón Artacho, F.J. Planes, E. Stalidzans, A. Maass, S. Vempala, M. Hucka, M.A. Saunders, C.D. Maranas, N.E. Lewis, T. Sauter, B. Palsson, I. Thiele, R.M.T. Fleming, Creation and analysis of biochemical constraint-based models using the COBRA Toolbox v.3.0, Nature Protocols 14 (2019) 639–702. https://doi.org/10.1038/s41596-018-0098-2.

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

[39] R.H.S. Diniz, J.C. Villada, M.C.T. Alvim, P.M.P. Vidigal, N.M. Vieira, M. Lamas-Maceiras, M.E. Cerdán, M.I. González-Siso, P.J. Lahtvee, W.B. da Silveira, Transcriptome analysis of the thermotolerant yeast Kluyveromyces marxianus CCT 7735 under ethanol stress, Applied Microbiology and Biotechnology 101 (2017) 6969–6980. https://doi.org/10.1007/s00253-017-8432-0.

[40] W. Mo, M. Wang, R. Zhan, Y. Yu, Y. He, H. Lu, Kluyveromyces marxianus developing ethanol tolerance during adaptive evolution with significant improvements

of multiple pathways, Biotechnology for Biofuels 12 (2019) 1–15. https://doi.org/10.1186/s13068-019-1393-z.

[41] B. Langmead, S.L. Salzberg, Fast gapped-read alignment with Bowtie 2, Nat Methods 9 (2012) 357–359. https://doi.org/10.1038/nmeth.1923.

[42] Y. Liao, G.K. Smyth, W. Shi, featureCounts: an efficient general purpose program for assigning sequence reads to genomic features, Bioinformatics 30 (2014) 923–930. https://doi.org/10.1093/bioinformatics/btt656.

[43] M.I. Love, W. Huber, S. Anders, Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2, Genome Biol 15 (2014) 1–21. https://doi.org/10.1186/s13059-014-0550-8.

[44] F. Almeida-Silva, T.M. Venancio, BioNERO: an all-in-one R/Bioconductor package for comprehensive and easy biological network reconstruction, Funct Integr Genomics 22 (2022) 131–136. https://doi.org/10.1007/s10142-021-00821-9.

[45] R.C. Gentleman, V.J. Carey, D.M. Bates, B. Bolstad, M. Dettling, S. Dudoit, B. Ellis, L. Gautier, Y. Ge, J. Gentry, K. Hornik, T. Hothorn, W. Huber, S. Iacus, R. Irizarry, F. Leisch, C. Li, M. Maechler, A.J. Rossini, G. Sawitzki, C. Smith, G. Smyth, L. Tierney, J.Y. Yang, J. Zhang, Bioconductor: open software development for computational biology and bioinformatics, Genome Biol 5 (2004) 1–16. https://doi.org/10.1186/gb-2004-5-10-r80.

[46] S. Carbon, E. Douglass, B.M. Good, D.R. Unni, N.L. Harris, C.J. Mungall, S. Basu, R.L. Chisholm, R.J. Dodson, E. Hartline, P. Fey, P.D. Thomas, L.P. Albou, D. Ebert, M.J. Kesling, H. Mi, A. Muruganujan, X. Huang, T. Mushayahama, S.A. LaBonte, D.A. Siegele, G. Antonazzo, H. Attrill, N.H. Brown, P. Garapati, S.J. Marygold, V. Trovisco, G. dos Santos, K. Falls, C. Tabone, P. Zhou, J.L. Goodman, V.B. Strelets, J. Thurmond, P. Garmiri, R. Ishtiaq, M. Rodríguez-López, M.L.

Acencio, M. Kuiper, A. Lægreid, C. Logie, R.C. Lovering, B. Kramarz, S.C.C. Saverimuttu, S.M. Pinheiro, H. Gunn, R. Su, K.E. Thurlow, M. Chibucos, M. Giglio, S. Nadendla, J. Munro, R. Jackson, M.J. Duesbury, N. Del-Toro, B.H.M. Meldal, K. Paneerselvam, L. Perfetto, P. Porras, S. Orchard, A. Shrivastava, H.Y. Chang, R.D. Finn, A.L. Mitchell, N.D. Rawlings, L. Richardson, A. Sangrador-Vegas, J.A. Blake, K.R. Christie, M.E. Dolan, H.J. Drabkin, D.P. Hill, L. Ni, D.M. Sitnikov, M.A. Harris, S.G. Oliver, K. Rutherford, V. Wood, J. Hayles, J. Bähler, E.R. Bolton, J.L. de Pons, M.R. Dwinell, G.T. Hayman, M.L. Kaldunski, A.E. Kwitek, S.J.F. Laulederkind, C. Plasterer, M.A. Tutaj, M. Vedi, S.J. Wang, P. D'Eustachio, L. Matthews, J.P. Balhoff, S.A. Aleksander, M.J. Alexander, J.M. Cherry, S.R. Engel, F. Gondwe, K. Karra, S.R. Miyasato, R.S. Nash, M. Simison, M.S. Skrzypek, S. Weng, E.D. Wong, M. Feuermann, P. Gaudet, A. Morgat, E. Bakker, T.Z. Berardini, L. Reiser, S. Subramaniam, E. Huala, C.N. Arighi, A. Auchincloss, K. Axelsen, G. Argoud-Puy, A. Bateman, M.C. Blatter, E. Boutet, E. Bowler, L. Breuza, A. Bridge, R. Britto, H. Bye-A-Jee, C.C. Casas, E. Coudert, P. Denny, A. Es-Treicher, M.L. Famiglietti, G. Georghiou, A.N. Gos, N. Gruaz-Gumowski, E. Hatton-Ellis, C. Hulo, A. Ignatchenko, F. Jungo, K. Laiho, P. Le Mercier, D. Lieberherr, A. Lock, Y. Lussi, A. MacDougall, M. Ma-Grane, M.J. Martin, P. Masson, D.A. Natale, N. Hyka-Nouspikel, S. Orchard, I. Pedruzzi, L. Pourcel, S. Poux, S. Pundir, C. Rivoire, E. Speretta, S. Sundaram, N. Tyagi, K. Warner, R. Zaru, C.H. Wu, A.D. Diehl, J.N. Chan, C. Grove, R.Y.N. Lee, H.M. Muller, D. Raciti, K. van Auken, P.W. Sternberg, M. Berriman, M. Paulini, K. Howe, S. Gao, A. Wright, L. Stein, D.G. Howe, S. Toro, M. Westerfield, P. Jaiswal, L. Cooper, J. Elser, The Gene Ontology resource: enriching a GOld mine, Nucleic Acids Research 49 (2021) D325–D334. https://doi.org/10.1093/NAR/GKAA1113.

[47] P. Jones, D. Binns, H.-Y. Chang, M. Fraser, W. Li, C. McAnulla, H. McWilliam, J. Maslen, A. Mitchell, G. Nuka, S. Pesseat, A.F. Quinn, A. Sangrador-Vegas, M. Scheremetjew, S.-Y. Yong, R. Lopez, S. Hunter, InterProScan 5: genome-scale protein function classification., Bioinformatics (Oxford, England) 30 (2014) 1236–40. https://doi.org/10.1093/bioinformatics/btu031.

[48] M.C. Teixeira, R. Viana, M. Palma, J. Oliveira, M. Galocha, M.N. Mota, D. Couceiro, M.G. Pereira, M. Antunes, I.V. Costa, P. Pais, C. Parada, C. Chaouiya, I. Sá-Correia, P.T. Monteiro, YEASTRACT+: a portal for the exploitation of global transcription regulation and metabolic model data in yeast biotechnology and pathogenesis, Nucleic Acids Research 51 (2023) D785–D791. https://doi.org/10.1093/nar/gkac1041.

[49] S. Chandrasekaran, N.D. Price, Probabilistic integrative modeling of genome-scale metabolic and regulatory networks in Escherichia coli and Mycobacterium tuberculosis, Proceedings of the National Academy of Sciences of the United States of America 107 (2010) 17845–17850. https://doi.org/10.1073/pnas.1005139107.

[50] M.L. Jenior, T.J.M. Jr, B.V. Dougherty, J.A. Papin, Transcriptome-guided parsimonious flux analysis improves predictions with metabolic networks in complex environments, PLOS Computational Biology 16 (2020) e1007099. https://doi.org/10.1371/journal.pcbi.1007099.

[51] Q.K. Beg, A. Vazquez, J. Ernst, M.A. De Menezes, Z. Bar-Joseph, A.L. Barabási, Z.N. Oltvai, Intracellular crowding defines the mode and sequence of substrate uptake by Escherichia coli and constrains its metabolic activity, Proceedings of the National Academy of Sciences of the United States of America 104 (2007) 12663–12668. https://doi.org/10.1073/pnas.0609845104.

[52] R.H.S. Diniz, W.B. Silveira, L.G. Fietto, F.M.L. Passos, The high fermentative metabolism of Kluyveromyces marxianus UFV-3 relies on the increased expression of key lactose metabolic enzymes, Antonie van Leeuwenhoek, International Journal of General and Molecular Microbiology 101 (2012) 541–550. https://doi.org/10.1007/s10482-011-9668-9.

[53] R.H.S. Diniz, M.Q.R.B. Rodrigues, L.G. Fietto, F.M.L. Passos, W.B. Silveira, Optimizing and validating the production of ethanol from cheese whey permeate by Kluyveromyces marxianus UFV-3, Biocatalysis and Agricultural Biotechnology 3 (2014) 111–117. https://doi.org/10.1016/j.bcab.2013.09.002.

[54] P.G. Ferreira, F.A. da Silveira, R.C.V. dos Santos, H.L.A. Genier, R.H.S. Diniz, J.I. Ribeiro, L.G. Fietto, F.M.L. Passos, W.B. da Silveira, Optimizing ethanol production by thermotolerant Kluyveromyces marxianus CCT 7735 in a mixture of sugarcane bagasse and ricotta whey, Food Science and Biotechnology 24 (2015) 1421–1427. https://doi.org/10.1007/s10068-015-0182-0.

[55] L. Signori, S. Passolunghi, L. Ruohonen, D. Porro, P. Branduardi, Effect of oxygenation and temperature on glucose-xylose fermentation in Kluyveromyces marxianus CBS712 strain, Microb Cell Fact 13 (2014) 51. https://doi.org/10.1186/1475-2859-13-51.

[56] F.A. Silveira, D.L. de Oliveira Soares, K.W. Bang, T.R. Balbino, M.A. de Moura Ferreira, R.H.S. Diniz, L.A. de Lima, M.M. Brandão, S.G. Villas-Bôas, W.B. da Silveira, Assessment of ethanol tolerance of Kluyveromyces marxianus CCT 7735 selected by adaptive laboratory evolution, Applied Microbiology and Biotechnology 104 (2020) 7483–7494. https://doi.org/10.1007/s00253-020-10768-9.

3 INVESTIGATING PROTEIN ALLOCATION: REDISTRIBUTION; *IN VIVO* USAGE, AND PROMISCUITY

3.1 Protein constraints in genome-scale metabolic models: Data integration, parameter estimation, and prediction of metabolic phenotypes

Text adapted from the unmarked revised manuscript as accepted by the Biotechnology and Bioengineering journal, available at: https://doi.org/10.1002/bit.28650.

Abstract

Genome-scale metabolic models provide a valuable resource to study metabolism and cell physiology. These models are employed with approaches from the constraint-based modelling framework to predict metabolic and physiological phenotypes. The prediction performance of genome-scale metabolic models can be improved by including protein constraints. The resulting protein-constrained models consider data on turnover numbers (k_{cat}) and facilitate the integration of protein abundances. In this systematic review, we present and discuss the current state-of-the-art regarding the estimation of kinetic parameters used in protein-constrained models. We also highlight how data-driven and constraint-based approaches can aid the estimation of turnover numbers and their usage in improving predictions of cellular phenotypes. Lastly, we identify standing challenges in protein-constrained metabolic models and provide a perspective regarding future approaches to improve the predictive performance.

Keywords: catalytic rates; enzyme concentration; parameterisation; machine learning; flux balance analysis.

Introduction

One of the aspirations of biological systems modelling is to identify organizing principles underlying molecular functions and to understand how the multitude of processes of cellular physiology gives rise to diverse phenotypes (Palsson, 2015). Constraint-based approaches have been widely used to simulate and predict phenotypes based on genome-scale reconstructions of metabolic networks. These reconstructions, termed genome-scale metabolic models (GEMs), comprise collections of metabolic reactions and their associated metabolites available in metabolic pathway databases, such as: KEGG, MetaCyc, BiGG, BRENDA and SABIO-RK, obtained by using annotated genomes (Thiele and Palsson, 2010). As a result, GEMs include the correspondence between genes, their respective proteins and the reactions that these proteins catalyse, captured in gene-protein-reaction (GPR) rules. These GPR rules provide a direct link between the genotype and molecular phenotype on the level of reaction rates (i.e., fluxes) (Machado et al., 2016). Constraint-based approaches predict metabolic flux distributions by optimizing one or more objective functions under a set of physicochemical constraints. A key, first representative of these approaches is Flux Balance Analysis (FBA) (Orth et al., 2010), which uses linear optimization to calculate metabolic flux based on two assumptions: (i) the network is at steady-state, whereby there is no change in the concentration of intracellular metabolites and (ii) cells evolved to optimize a metabolic objective, such as growth or production of an important metabolite (Gianchandani et al., 2010). The first assumption results in an under-determined system of linear equalities arising from mass balance equations, with fluxes as unknowns. By imposing constraints that represent cellular growth conditions, the solution space can be restricted, resulting in more biologically relevant predictions. These constraints include the upper and lower bounds for metabolic fluxes, (ir)reversibility of reactions and exchanges with the environment. The second assumption further carves a part of the feasible space, resulting from the application of the constraints, and narrows down the physiologically relevant solutions.

While conventional GEMs provide a useful framework for studying metabolism in a large-scale, there are still many phenotypes that elude their predictive capabilities. More specifically, metabolic shifts that occur at higher growth rates cannot be accurately simulated by conventional GEMs, with notable examples including the overflow metabolism in bacteria (Basan et al., 2015), the Crabtree effect in

Saccharomyces cerevisiae (Sánchez et al., 2017), and the Warburg effect in human cancer cells (Shlomi et al., 2011) (see Section 3.1 for detailed elaboration). Phenotype predictions of GEMs can be improved by integrating protein constraints related to enzyme kinetic parameters and enzyme concentrations. The turnover number, k_{cat}, of an enzyme is a key kinetic parameter that refers to a first-order rate constant with the unit of s^{-1} that describes the conversion of a substrate to product per unit of time, as accelerated by the enzyme. However, despite the advantages and advances in integrating enzyme parameters in GEMs, obtaining k_{cat} values remains challenging. Measurements of the kcatome, a subset of the kinetome that includes the turnover numbers of all enzymes, depends on purification of specific enzymes, which often is difficult (Nilsson et al., 2017). Further, there is a lack of knowledge of the cofactors and co-enzymes required for enzymatic function, which hinders *in vitro* measurements of k_{cat} values (Davidi et al., 2016). Even when *in vitro* measurements are available, the usage of the resulting values in pcGEMs is challenging since the kinetic data is obtained from non-physiological conditions, which becomes a source for a discrepancy with real physiological conditions (Chen and Nielsen, 2021).

Enzyme concentrations are the additional constraint added to GEMs along with turnover numbers. They are derived from absolute protein abundances, usually obtained from quantitative proteomics experiments (Lahtvee et al., 2017; Sánchez et al., 2017). However, like estimates of turnover numbers, absolute proteomics measurements are still difficult to obtain. There are several challenges, including: (i) a large portion of proteins is still undetected due to limitations of current mass spectrometry techniques (Pappireddi et al., 2019), (ii) the ionization efficiency is heavily affected by the protein's physicochemical properties (Otto et al., 2014), (iii) high cost of equipment and reagents (Swiatly et al., 2018), (iv) lack of a standardized approach for measuring absolute abundances (Calderón-Celis et al., 2018). Methods for absolute protein quantification, such as isobaric tagging, stable isotope labelling and others have been reviewed by (Lindemann et al., 2017). In addition, the reproducibility of absolute protein quantification in different samples is often inconsistent (Millán-Oropeza et al., 2022). In non-model species, proteomics studies are further complicated by the lack of physiological and genomics information, especially annotated genomes (Heck and Neely, 2020). Nevertheless, there have been many

efforts in using computational approaches to estimate k_{cat} values and protein abundance, which is explored in the further sections of this review.

This review focuses on protein-constrained GEMs (pcGEMs), the estimation of the parameters they comprise, and their usage in predicting phenotypes (Figure 1). First, we present the current methods of integration of k_{cat} values in GEMs, providing a summary of approaches that use pcGEMs, and the phenotypes that can be determined by having access to k_{cat} and enzyme abundance values. Next, we discuss the approaches for estimation and correction of k_{cat} values. We then examine approaches for prediction of protein abundance and their usage in pcGEMs. Lastly, we identify and discuss current challenges in parameter estimation and their integration in pcGEMs, and provide a perspective regarding future approaches to improve the predictive performance values.

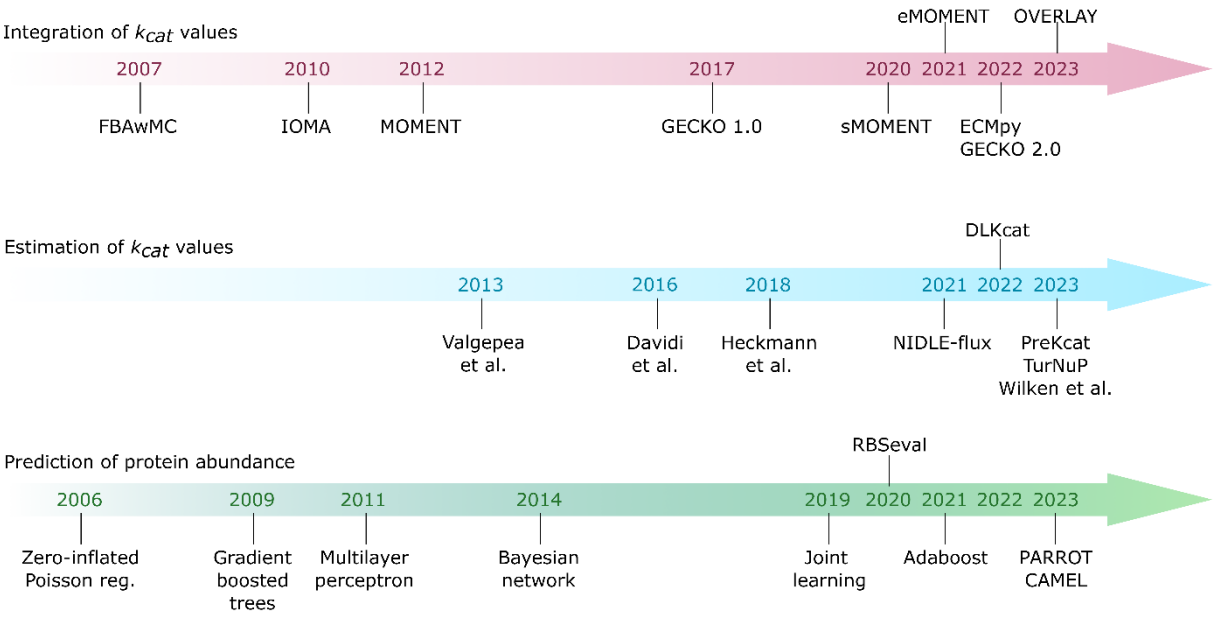

Figure 1. Timelines for approaches related to protein-constrained GEMs. Three timelines are considered regarding the following problems addressed in protein-constrained GEMs: (i) integration of turnover numbers in GEMs, (ii) estimation of turnover numbers, and (iii) prediction of protein abundances.

Protein-constrained genome-scale metabolic models

The shortcomings of conventional GEMs have propelled the development of approaches to improve their prediction capabilities (Ye et al., 2022). The development and usage of pcGEMs represents one means to overcome these limitations by

considering the properties of enzymes that determine reaction fluxes. These pcGEMs inherit all constraints from conventional GEMs, such that:

$$\mathbf{S} \cdot \mathbf{v} = 0,$$

where **S** is the stoichiometric matrix and **v** is the flux distribution vector; and:

$$\mathbf{v_{min}} \leq \mathbf{v} \leq \mathbf{v_{max}},$$

where $\mathbf{v_{min}}$ and $\mathbf{v_{max}}$ are the lower and upper bounds for metabolic flux, respectively. This allows for classic simulation approaches, like flux balance analysis (FBA) (Orth et al., 2010), to also be applicable with pcGEMs. A key difference is the inclusion of the constraint, common to most approaches:

$$v_j \leq k_{cat}^{ij} \cdot [E_i],$$

where v_j is the metabolic flux of the reaction j, $[E_i]$ is the internal concentration of enzyme i, and k_{cat}^{ij} is the enzyme turnover number of reaction j catalysed by an enzyme i (Adadi et al., 2012; Sánchez et al., 2017).

The constraints applied to pcGEMs are straightforward when considering reactions catalysed by a single enzyme. There are, however, other associations formulated in GPR rules, such as isozymes and enzyme complexes; in addition, a gene can participate in multiple GPR rules due to enzyme promiscuity (Amin et al., 2019). Enzyme complexes can be divided into homo- or heteromeric. Homomeric enzymes are complexes composed of identical subunits, and enzymatic constraints can directly be defined for such complexes. Heteromeric enzymes, however, are complexes composed of different protein subunits encoded by different genes. This makes the usage of k_{cat} values challenging, since it can become unclear which subunits contain active sites (Davidi et al., 2016).

Besides pcGEMs, other notable developments to improve conventional GEMs beyond protein allocation is to consider the whole machinery involved in protein biosynthesis and other cellular processes. These come in two flavours – models of metabolism and macromolecular expression (ME-models) and models of resource balance analysis (RBA) (Figure 2). In the former, the processes of transcription and translation are represented in the model, and metabolic fluxes are predicted by including substrate-enzyme binding and product-enzyme dissociation reactions. They may also include information on protein translocation, compartmentalization, folding and thermostability (Lloyd et al., 2018). In the latter, macromolecular processes such

as secretion and protein folding through chaperones are also taken into account along with protein biosynthesis (Goelzer et al., 2015). These approaches have been recently reviewed by De Becker et al. (2022); Kerkhoven (2022); and Regueira et al. (2021).

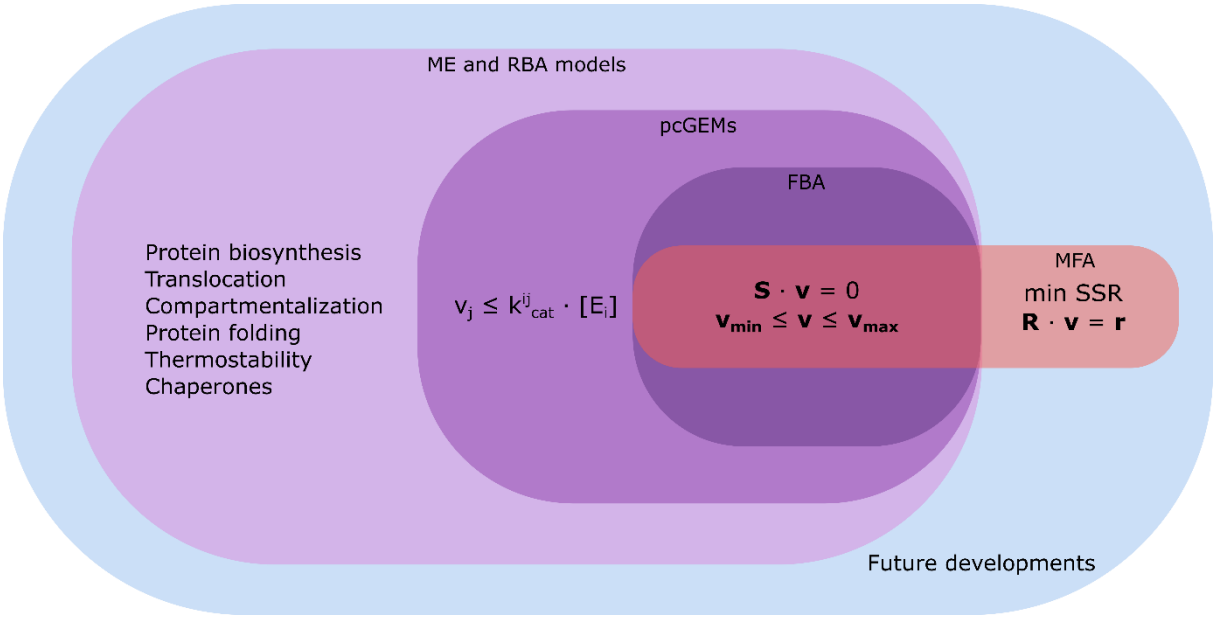

Figure 2. Relationship between different types of GEMs and approaches for their analyses. As pcGEMs improve on conventional GEMs, they embed the constraints from FBA. The ME and RBA models, likewise, envelop pcGEMs by also considering enzyme catalytic rates and enzyme abundances, alongside the additional protein biosynthesis machinery. MFA denotes approaches that rely on stoichiometry in addition to atom mappings to estimate fluxes based on data on labelling patterns.

The specific way in which enzyme constraints are encoded varies with each approach. Two groups of approaches can be identified based on whether turnover numbers and enzyme mass balance are explicitly represented in the stoichiometric matrix. The first group comprises FBAwMC, MOMENT, eMOMENT, and ECMpy, which do not change the original stoichiometric matrix, while the second includes GECKO, sMOMENT, PAM, and OVERLAY, which expand the stoichiometric matrix in encoding the protein constraints. The definition and specifics of each approach are discussed in detail in the Section 3, below. If proteomics measurements are available, they can be used to specify the concentration of enzymes $[E_i]$. In the case when such measurements are not available, a common assumption is that enzyme usage is limited according to the total enzyme pool E_{pool}, derived from the total protein content,

denoting the total usage of all metabolic enzymes integrated in the model (Bekiaris and Klamt, 2020; Sánchez et al., 2017). Another feature of pcGEMs is the splitting of each reversible reaction into two irreversible forward and backward reaction. This way, only positive flux values are calculated, while also making it possible to associate different k_{cat} values for each reaction, since substrate affinity could be different (Adadi et al., 2012; Davidi et al., 2016).

Integration of turnover numbers in GEMs

Over the years, many methods have been developed to integrate enzymatic information in GEMs. The first method to consider enzyme parameters was FBA with Molecular Crowding (FBAwMC), which uses information about the solvent capacity for attainable enzyme concentrations inside the cytoplasm (Beg et al., 2007). The optimization problem solved by FBAwMC is to maximize the growth rate subject to FBA constraints plus limitations in the crowding coefficient a. This is attained by constraints that involve the cell volume V such that:

$$\sum_{i=1}^{N} v_i \, n_i \leq V,$$

where v_i is the molar volume of enzyme i and n_i are the level of enzymes. This equation is reformulated by dividing by the cell mass M, which gives that:

$$\sum_{i=1}^{N} v_i \, E_i \leq \frac{1}{C},$$

where $E_i = n_i/M$, the enzyme concentrations; and $C = M/V$, the cytoplasmic density. By considering that the flux distribution v_i is defined as:

$$v_i = b_i E_i,$$

where $b_i = x_i k_i$, with x_i being the concentration of substrates, products, activators and inhibitors associated with the reaction i; and k_i being the k_{cat} for an enzyme i; the constraint applied to simulations is then defined as:

$$\sum_{i=1}^{N} a_i \, v_i \leq 1,$$

such that the crowding coefficient is $a_i = C v_i / b_i$.

Another method, termed IOMA (Integrative Omics Metabolic Analysis) considered enzyme turnover numbers along with proteomics and metabolomics data (Yizhak et al., 2010). IOMA relies on a Michaelis-Menten-like rate equation to estimate flux distributions. It considers relative protein levels and enzyme kinetics information

such as k_{cat} values and V_{max}. In this way, the following constraint is added to the classic FBA constraints:

$$v = \frac{e}{e^{ref}}(a^{+}v_{max}^{+} - a^{-}v_{max}^{-}),$$

where *e* is the concentration of enzymes, e_{ref} the concentration of enzymes at a reference condition, a^{+} and a^{-} the saturation values for enzymes in forward and reverse reactions, respectively, and V_{max}^{+} and V_{max}^{-} the maximal flux for forward and reverse reactions, respectively. The saturation values for enzymes are calculated as:

$$a^{+} = \prod_{s_i \in s}\left(\frac{s_i}{k_m, s_i + s_i}\right), \quad a^{-} = \prod_{p_i \in p}\left(\frac{p_i}{k_m, p_i + p_i}\right),$$

where s_i and p_i are respectively the concentrations of substrates and products i, and k_m, s_i and k_m, p_i are the dissociation constants for substrates and products, respectively.

These previously described methods provided an improvement over FBA, but as pointed out by Adadi et al. (2012), they depended on utilizing experimentally determined uptake rates to predict phenotypes across different conditions. To address this issue, Adadi et al. (2012) developed MetabOlic Modeling with Enzyme kineTics (MOMENT), which expands the FBAwMC approach by taking into account the maximal cellular capacity of enzymes. The MOMENT approach also improved on the handling of isozymes and enzyme complexes. A reaction catalysed by a single enzyme is constrained by the equation $v_j \leq k_{cat}^{ij} \cdot [E_i]$. However, for reactions that can be catalysed by two enzymes *a* or *b*, the equation changes to:

$$v_j \leq k_{cat}^{ij} \cdot [E_a + E_b].$$

For reactions catalysed by enzyme complexes, the formulation changes to:

$$v_j \leq k_{cat}^{ij} \cdot \min[E_a, E_b].$$

Similar to FBAwMC, MOMENT also uses a constraint on the enzyme solvent capacity:

$$\sum[E_i] \cdot MW_i \leq C\left[\frac{g_{protein}}{g_{DW}}\right],$$

where MW_i is the molecular weight of protein i, C is the total weight of proteins.

Improvements and derivations of the MOMENT approach have been developed recently. Bekiaris & Klamt (2020) have extended and simplified MOMENT, resulting in sMOMENT (short MOMENT). In this approach, the same constraints of MOMENT are used, but resulting in a model with significantly fewer variables while yielding the same

results as MOMENT. This is achieved by reformulating the enzyme solvent capacity constraint, such that:

$$\sum v_i \cdot \frac{MW_i}{k_{cat},i} \leq \sum [E_i] \cdot MW_i \leq P,$$

where P is the threshold (g/g_{DW}) of enzymes covered by the pcGEM. This equation is further reformulated as:

$$-\sum v_i \cdot \frac{MW_i}{k_{cat,i}} + v_{pool} = 0; \;\; v_{pool} \leq P,$$

where v_{pool} represents the mass of all enzymes in the pcGEM needed to catalyse the reactions in the model. The sMOMENT approach also changes how enzyme complexes are used by the model. By considering the enzyme costs c_i of a reaction i, such that:

$$c_i = \frac{MW_i}{k_{cat,i}},$$

then for all enzymes catalysing reaction $i,$ or all sub-units of an enzyme complex, the minimum value is used:

$$c_i = min\left(\left\{\frac{MW_{e^1}}{k_{cat,e^1}}, \frac{MW_{e^2}}{k_{cat,e^2}}, \ldots\right\}; E^{i1}, E^{i2}, \ldots \in E(i).\right.$$

Another extension to MOMENT is the approach termed eMOMENT, which introduces enzyme promiscuity as a new constraint to the base MOMENT formulation (Wendering and Nikoloski, 2022). This constraint is defined as:

$$\sum_{i \in GPR_k} E^r_{k,i} = E^g_k, \quad \forall k \in G,$$

where G is the set of genes in the model, E_k is the abundance of enzyme k, and r is the set of reactions in the model.

An approach inspired by the reformulations and reduction of model complexity introduced by sMOMENT has been developed by Mao et al. (2022), termed ECMpy. This approach introduces enzyme constraints in the model without adding new reactions or explicitly accounting for enzymes in the stoichiometric matrix. Instead, it uses a single constraint, defined as:

$$\sum_{i=1}^{n} \frac{v_i \cdot MW_i}{\sigma_i \cdot k_{cat,i}} \leq p_{tot} \cdot f,$$

where σ is the saturation coefficient of the enzyme i, p_{tot} is the total protein content in the model, and f is mass fraction of enzymes.

An approach similar to MOMENT, termed GEM with Enzymatic Constraints using Kinetic and Omics data (GECKO), also integrates k_{cat} values to limit metabolic

flux according to its maximum capacity (Sánchez et al., 2017). GECKO also limits reactions according to the abundance of each enzyme present in the model, given that absolute proteomics measurements are available. The GECKO approach uses the same constraint as MOMENT for reactions catalysed by single enzymes, $v_j \leq k_{cat}^{ij} \cdot [E_i]$. It differs in its handling of isozymes, promiscuous enzymes (not considered in MOMENT), and enzyme complexes. For isozymes, the flux is constrained as:

$$v_j \leq \sum_i k_{cat}^{ij} \cdot [E_i].$$

In this scenario, the reaction is split into as many different reactions as there are isozymes capable of catalysing that reaction, each with only one enzyme. Then, an intermediate reaction termed “arm reaction” is added to keep the original upper bound of the reaction, using a pseudo-metabolite representing an intermediate state between a substrate and a product. This allows for the enzyme of each reaction to be assigned a different k_{cat} value, as substrate affinity could be different. For reactions catalysed by promiscuous enzymes:

$$\sum_j \frac{v_j}{k_{cat}^{ij}} \leq [E_i].$$

This arrangement also allows considering different k_{cat} values for each reaction catalysed by the same enzyme. Additionally, these reactions share the same upper bound of enzyme availability. Lastly, for reactions catalysed by enzyme complexes, the stoichiometry s_{ik} of the enzyme subunit U_{ik} and its concentration $[U_{ik}]$ are considered:

$$v_j \leq k_{cat}^{ij} \cdot \min_k \left(\frac{[U_{ik}]}{s_{ik}}\right).$$

The GECKO approach has also been improved upon. A method developed by Alter et al. (2021), termed Protein Allocation Model (PAM), uses the base formulation of GECKO for integrating k_{cat} values and proteomics data, and reimplements the idea of protein allocation sectors as developed by Mori et al. (2016) in the approach termed Constraint Allocation FBA (CAFBA). In CAFBA, the total enzyme usage of the model is divided into four sectors: ribosomal proteins, biosynthetic proteins, uptake and transport proteins, and housekeeping proteins. The PAM approach instead divides the total enzyme usage into three sections: translational proteins, which are enzymes related to protein biosynthesis and is directly associated with the growth rate; unused enzymes, defined as those enzymes that exists in an over-abundance state, where more proteins were produced than it was necessary for the cell during a certain

physiological state, and reduces the growth rate; and active enzymes, those actively involved in catalysing reactions in metabolism. The translational protein sector Φ_T, is defined as:

$$\Phi_T = \Phi_{T,0} + w_T \mu \ ,$$

where $\Phi_{T,0}$ is the measure of translational enzyme concentration at zero growth, w_T is the maximum ribosomal elongation rate, and μ is the growth rate. Regarding the unused enzyme sector Φ_{UE}, it is expressed as:

$$\Phi_{UE} = \Phi_{UE,0} - w_{UE} v_s,$$

where $\Phi_{UE,0}$ is the measure of unused enzyme concentration at zero growth, w_{UE} is the measure of the increase in enzyme usage efficiency, and v_s is the substrate uptake rate. With respect to the active enzyme sector Φ_{AE}, it has to account for enzyme mass balances and k_{cat} values of all enzymes that catalyse metabolic reactions. It is defined as:

$$\Phi_{AE} = \sum_e^E v_e \frac{M_e}{k_{cat,e}},$$

where v_e is the flux of a reaction catalysed by the enzyme e, and M_e is the molar mass of the enzyme. Thus, the total enzyme mass concentration of the model can be described as the sum of all protein sections:

$$\Phi_{P,c} = \Phi_T + \Phi_{AE} + \Phi_{UE}.$$

An approach, termed OVERLAY, proposes a different formulation (Yao et al., 2023). It integrates catalytic rates in the form of the effective turnover rate, k_{eff}. In contrast to other approaches, it considers enzyme complexes separately from other enzymes, treating complexes as a single entity in the model. Further, for each reaction catalysed by an enzyme, OVERLAY adds a pair of forward and reverse enzymes to account for reversible reactions, while spontaneous reactions are ignored. In terms of model constraints, the k_{eff} values are used to define the lower and upper bounds of a reaction v, such that:

$$-k_{eff}^{avg} Ie_{rev} \leq v \leq k_{eff}^{avg} Ie_{for},$$

where k_{eff}^{avg} is defined as a basal value of k_{eff} (assumed to be 65s^{-1}), e_{rev} is enzyme concentration for the reverse reaction, and e_{for} is the enzyme concentration for the forward reaction.

The described approaches vary widely in how they are implemented, how much of the reconstruction steps are automated, and how easy is to peruse their documentation. Focusing on the approaches that are available in public repositories (e.g., GitHub), the sMOMENT approach is bundled with a workflow termed AutoPACMEN, a Python package that allows for automated reconstruction of pcGEMs. A step-by-step tutorial is included in the supplementary information of the original manuscript, with a more detailed documentation provided in the Python package manual. The ECMpy approach is also implemented in Python, with its main functions contained in a single script. A step-by-step guide is available in the form of Jupyter notebooks, which reproduces the reconstruction of eciML1515 as performed in its manuscript. The GECKO approach, on the other hand, is available as a MATLAB package. It supports the reconstruction and refinement of pcGEMs, and the integration of proteomics measurements. The documentation for the GECKO approach is not extensive, but some information is included as comments in its main functions. The GECKO approach, while mainly developed for MATLAB, also includes a Python package to integrate protein abundances and to allow interface with cobrapy (Ebrahim et al., 2013). It is important to highlight that since sMOMENT, ECMpy, and GECKO follow similar formulations, they are able to generate very similar models with equivalent phenotype prediction capabilities. GECKO models, however, are notoriously more complex than sMOMENT or ECMpy models, since GECKO explicitly introduces enzyme usage pseudoreactions and pseudometabolites to consider enzyme constraints. The PAM approach generates models similar to GECKO models, but with the addition of proteome sectors. It is also available as a MATLAB package. It currently lacks a detailed documentation on how to execute the tool, but an example code is provided. Lastly, the OVERLAY approach is developed for MATLAB, and includes documentation and a step-by-step tutorial to reproduce the pcGEM generated in its manuscript. It provides an automated tool to reconstruct pcGEMs, with model complexity comparable to sMOMENT and ECMpy.

After enzyme constraints are integrated into GEMs, giving rise to pcGEMs, a plethora of previously unattainable phenotypes are now able to be simulated. Simulations of metabolic switches, like the overflow metabolism, are often missed by conventional GEMs. Overflow metabolism denotes the phenomenon where metabolic flux is totally or partially redirected from respiratory pathways to fermentation

pathways, despite the availability of oxygen (de Alteriis et al., 2018). This phenomenon is also known as the Crabtree effect and the Warburg effect in the context of yeast and human cancer cells, respectively, and occurs when the carbon source availability exceeds the capability of an organism in assimilating it (Li et al., 2022).

Phenotypes that can be determined having access to k_{cat} values

A metabolic phenomenon often missed by conventional GEMs is diauxic growth. In a setting where multiple carbon sources are available, conventional FBA predicts simultaneous uptake and usage of all carbon sources, which is biologically unrealistic. With the FBAwMC approach, the *E. coli* model MG1655 was simulated in a condition where five different carbon sources where available: glucose, galactose, maltose, glycerol and lactate. In this set-up, the sequence of substrate uptake and consumption matched experimental data, with glucose being used first and exclusively, followed by galactose, lactate, maltose and glycerol (Beg et al., 2007). The Crabtree effect in *S. cerevisiae* was captured in the pcGEM ecYeast7 by simulating a glucose-limited chemostat with increasing growth rates. As the growth rate increased, there was a linear increase for the uptake of glucose, O_2, and production of CO_2. At a growth rate of 0.3 h^{-1}, the uptake of glucose and production of ethanol sharply spiked, while O_2 consumption sharply decreased. However, the conventional GEM Yeast7 still predicted a linear increase of the uptake of glucose, O_2, and production of CO_2 (see the Figure 3A from the Sanchez et al. (2017) reference) In terms of pathway usage, the pcGEM predicted an increase in metabolic flux through glycolysis, while the flux through oxidative phosphorylation decreased (Sánchez et al., 2017).

Using the GECKO approach and dynamic FBA (dFBA) (Mahadevan et al., 2002), the *S. cerevisiae* model ecYeast8 can also predict the order of consumption of the carbon sources with a good correlation with experimental data. The dFBA method introduces kinetic equations for extracellular metabolites and biomass, allowing for a time-dependent simulation of metabolism. When growth is simulated using a combination of glucose and sucrose as carbon sources, ecYeast8 first predicts the consumption of glucose as the initial carbon source.When glucose is depleted, the hydrolysis of sucrose occurs with the subsequent consumption of glucose as carbon source. In this scenario, fructose was left unused until glucose was depleted, which was then used to support a third phase of growth (Moreno-Paz et al., 2022).

Considering the importance of k_{cat} values to simulate phenotypes such as diauxic growth and overflow metabolism, and the challenges with *in vitro* k_{cat} measurements, there is a growing need to find alternatives to experimental measurements, as we describe in the next section.

Approaches for estimation of k_{cat} values

Four different computational approaches have been proposed to estimate *in vivo* k_{cat} values. These approaches rely on the relationship between fluxes, enzyme concentration and k_{cat} values, or rely on data-driven models trained on enzyme biochemistry data and features derived from biological sequences.

The maximum rate of a reaction, V_{max}, can be determined by knowing the catalytic rate of the reaction when the enzyme is at their point of saturation and how much of that enzyme is present, given the relationship:

$$v_{max} = k_{cat} \cdot [E].$$

As this represents the maximum rate, effects of metabolites can only result in metabolic fluxes lower than V_{max}. This is accounted by a considering a function η of that captures the effect of metabolite concentrations and different parameters (e.g. equilibrium and Michaelis-Menten constants, K_{eq} and K_m). For an environmental condition C, the function η satisfies the expression:

$$0 \leq \eta\big(K, x(C)\big) \leq 1,$$

allowing for the metabolic flux to be expressed as:

$$v_j(C) \leq k_{cat}^{ij} \cdot [E_i(C)] \cdot \eta\big(K, x(C)\big).$$

If estimates of $v_j(C)$ and $[E_i(C)]$ are available, it is possible to calculate the apparent catalytic rate k_{app} by rearranging $v_j(C) = k_{app,j}(C) \cdot [E_j(C)]$ such that:

$$k_{app,j} = \frac{v_j(C)}{[E_i(C)]},$$

or by considering the relationship between k_{cat} and η*:*

$$k_{app,j} = k_{cat}^{ij} \cdot \eta(C),$$

which then leads to:

$$v_j(C) = k_{app,j} \cdot \eta(C).$$

This relation allows us to derive one of the three quantities if the other two are available. It was first used by Valgepea et al. (2013) to calculate k_{app} values by using

quantifications of the absolute proteome and metabolic flux analysis of *E. coli* cultivated in increasing growth rates. They calculated k_{app} values for 191 enzymes and found that as growth rate increases, there is a 3.7-fold increase in k_{app} values, which is discussed as a possible mechanism in which metabolic flux increases alongside growth rates.

Estimation of catalytic rates

The estimation of *in vivo* catalytic rates by Davidi et al. (2016) builds on the approach by Valgepea et al. (2013) and laid much of the groundwork in estimating the kcatome that following approaches later employed, by exploring the relationship $v_j(C) = [E(C)] \cdot k_{cat}^{ij} \cdot \eta(C)$. It was also demonstrated by Davidi et al. (2016) that these estimations could be used in pcGEMs. Proteomics measurements were used for $[E(C)]$, while $v_j(C)$ was determined by parsimonious enzyme FBA (pFBA), which minimizes the total flux through the network and constrains the model by growth rate and culture medium composition. By assessing proteomics experiments performed in 31 different growth conditions, they estimated condition-specific catalytic rates and took the maximum value as the maximum k_{app}, or k_{max}^{vivo}. The usage of predicted fluxes, instead of fluxes determined by metabolic flux analysis (Valgepea et al., 2011), allowed the estimation of catalytic rates for a number of enzymes larger than previously obtained by Valgepea et al. (2013). These estimated values were in good agreement with *in vitro* k_{cat} values, which shows the precision of the method in predicting catalytic rates.

Next, Heckmann et al. (2018) used machine learning to predict the catalytic rates k_{cat} and $k_{app,max}$ using a feature set composed of structural, biochemical and network data, such as molecular weight, structural disorder, active site structure and function, EC number, metabolic flux, *Km*, pH, and temperature. Five different regression models were trained and assessed: linear regression, partial least squares, elastic net, random forest, and a deep neural network. Based on the coefficient of determination (R^2) as a goodness-of-fit measure, the best performing model was the random forest, which achieved median R^2 scores of over 0.75, for both the training and test datasets. An analysis of feature importance revealed metabolic flux to be the most important feature for prediction of both catalytic rates. They found that models using $k_{app,max}$ values had better predictive capability than models using k_{cat} values, based

on root mean squared error (RMSE) of predicted enzyme usage values. A follow-up on the study used proteomics and fluxomics data to estimate the catalytic rates instead of using pFBA-predicted metabolic fluxes, achieving more precise estimations (Heckmann et al., 2020).

A further development using constraint-based approaches was performed by Xu et al. (2021). This approach rests on the observation that in both Davidi et al. (2016) and Heckmann et al. (2018) there are many expressed enzymes that carry no flux, called idle enzymes (Xu et al., 2021). For that reason, Xu et al. (2021) formulated a two-step mixed-integer linear program (MILP) termed NIDLE-flux, that maximizes the number of enzymes that carry flux. With this approach, it is then possible to increase the number of catalytic rate values that can be estimated. The resulting metabolic flux is then used together with protein abundance data to estimate k_{max}^{vivo} as in the approach of Davidi et al. (2016). This approach led to a 1.4-fold increase in the number of estimated k_{max}^{vivo} values, compared to estimations by Davidi et al, (2016) and Heckmann et al. (2020).

The next development on estimation of catalytic rates after NIDLE-flux is a novel constraint-based approach to estimate catalytic rates developed by Wilken et al. (2022). The approach relies on an objective function that minimizes the error function of predicted and measured fluxes and enzyme concentrations, such that:

$$L(\mathbf{k}_{\text{cat}}) = \min_{\text{v,e}} \frac{1}{|I|}\sum_{i\in I}\left(1-\frac{v_i}{\hat{v_i}}\right)^2 + \frac{1}{|J|}\sum_{j\in J}\left(1-\frac{e_j}{\hat{e_j}}\right)^2,$$

$$\textbf{s.t. Sv} = 0,$$

$$v_n = k_{cat,n}\cdot e_n \;\forall\; n \text{ metabolic reactions},$$

$$\sum_n e_n \leq E_{total},$$

where L is the error function, v_i is the flux through reaction i, and e_j is the concentration of the enzyme j. This optimization problem allows for the prediction of both metabolic fluxes and enzyme concentrations (also unknown variables in the function $L(\mathbf{k}_{\text{cat}})$, albeit not represented in the manuscript along with molecular weights of proteins), which given the relationship between these variables ($k_{cat} = \frac{v}{\hat{e}}$), it is possible to calculate the k_{cat} value that best fit the experimental data. Wilken et al. (2022) tested this optimization problem using experimental data from Heckmann et al. (2020), predicting metabolic fluxes and protein concentrations for a diverse range of growth

conditions. They found that using k_{cat} values estimated from this approach improves the accuracy of model predictions against experimental data by 35±2%. However, the approach in the present formulation includes quadratic constraints (due to $v_n = k_{cat,n} \cdot e_n$), which deserves further exploration of identifying the global optimum of the optimized function. Another data-driven method is the deep learning approach DLKcat, developed by Li et al. (2022). This method uses a combination of graph neural network, taking as input features derived from substrates in the SMILES format; and a convolutional neural network, using amino acid sequences as inputs. The networks also considered the substrate name, EC number, organism name and k_{cat} values. The DLKcat method thus requires significantly fewer inputs than the model developed by Heckmann et al. (2018). It was developed using the Python package PyTorch (Paszke et al., 2019). DLKcat was able to predict k_{cat} values of all enzymes of 343 fungal species, all of which were used for reconstruction of pcGEMs. A sequential Monte-Carlo-based approximate Bayesian computation was also used to correct *in vivo* k_{cat} estimations when these were significantly different from *in vitro* k_{cat} values. For assessing the predictions, RMSE values were calculated between experimental and predicted growth in *S. cerevisiae* and *Yarrowia lipolytica*, achieving lower values in each generation in the Bayesian approach training process and outperforming the original pcGEMs. The predictions obtained using DLKcat were made available to the public in an extensive database named GotEnzymes (Li et al., 2023), available at https://metabolicatlas.org/gotenzymes, containing over 25.7 million pairs of enzyme and substrates for over 8000 organisms. DLKcat itself is available in a GitHub repository and contains a step-by-step guide for the user to install the dependencies and run the trained models in a command line script.

While the DLKcat method proposes a powerful approach to predict k_{cat} values, some shortcomings were identified by other works; for instance, DLKcat accurately predicts only k_{cat} values of enzymes similar to those used in the training dataset, with decreasing accuracy for enzymes with more dissimilar amino acid sequences to those found in the training data (Kroll and Lercher, 2023). This has inspired the development of other data-based models using different algorithms and training data to improve predictions. In this regard, Yu et al., (2023) goes in a different direction compared to DLKcat, and proposes the usage of a pretrained language model to predict k_{cat} values

instead of a convolutional neural network. This approach, termed PreKcat, depends on amino acid sequences and the molecular structure of substrates, and addresses two problems: predicting K_m values and predicting k_{cat}/K_m ratios (i.e. catalytic efficiency). PreKcat achieved a Pearson correlation of 0.83 on the test dataset, which contained enzymes not seen in the training dataset. Similar to DLKcat, PreKcat was developed using PyTorch, and is available on a GitHub repository. The tool requires several pretrained language models preinstalled, and the documentation is still under construction as of writing.

Despite the improvements of PreKcat, the sequence similarity between enzymes in the training dataset and test dataset can still bias the predictions, even if the enzymes themselves are different. On this point, Kroll et al., (2023) proposed a new method, termed TurNuP, to predict k_{cat} values using a modified and re-trained Transformer Network. The model was trained using amino acid sequences, substrate and product IDs, reaction equations, and k_{cat} measurements, all collected from BRENDA, UniProt and Sabio-RK. The information was transformed into binary molecular fingerprints for each substrate and each product, in order to integrate the data. TurNuP was able to predict k_{cat} values with good agreement to experimental data, achieving a Pearson correlation of 0.67. For unseen reactions, which were not present in the original dataset, the performance was still good, achieving a Pearson correlation of 0.60. Kroll et al. (2023) also evaluated how sequence similarity between enzymes in the training and test datasets affects model performance. For enzymes with high sequence similarity (99-100%), the model achieves an R^2 score of 0.67, while for enzymes with low sequence similarity (0-40%), the R^2 score was 0.33. As done with previous tools, TurNuP was developed using PyTorch. It is available on a GitHub repository, along with all datasets deposited on Zenodo. Further, TurNuP is available in a web server (https://turnup.cs.hhu.de/), requiring no previous setup. It is important to highlight, however, that the described approaches were mostly focused on estimating k_{max}^{vivo} values for enzymes participating in reactions with simple GPR rules, e.g., a single enzyme catalysing a single reaction. For isozymes, however, it is difficult to determine k_{max}^{vivo} values as the kinetics of each enzyme might differ. This is tackled by (Davidi et al., 2016) by treating the isozymes as one single enzyme, using a sum of molecular weights of each isozyme to find the molecular weight of the lumped single enzyme. In contrast, Xu et al. (2021) uses the maximum abundance for

isozymes and the minimum abundance for enzyme complexes. Meanwhile, Heckmann et al. (2018) does not even consider complex GPR rules in their approach. This challenge is more pronounced for enzyme complexes, as to date few studies have described a possible way to assess their enzyme kinetics, with many focusing only on homomeric complexes. Homomeric complexes can be treated as a single element and thus assume only one catalytic rate value, but at the cost of forfeiting the specific mapping of enzyme data to proteins and their reactions. For heteromeric enzymes, Davidi et al. (2016) has proposed using the specific activity (SA) of the heteromeric complex instead of the k_{cat} value. The SA is defined as amount of product (in molar quantity) that is formed in a reaction per weight of enzyme per unit of time. It is calculated as:

$$SA_{app}[\mu mol \cdot mg^{-1} \cdot min^{-1}] = \frac{flux[\mu mol \cdot gCDW^{-1} \cdot min^{-1}]}{heteromeric\ complex\ mass\ fraction[mg \cdot gCDW^{-1}]}.$$

Correction of catalytic rates

The integration of k_{cat} and k_{app} values obtained from either biochemical assays or computational estimations can often lead to over-constrained models. This happens due to the inherent errors and uncertainties of these measurements. In the first iteration of the GECKO approach, k_{cat} values retrieved from the BRENDA database were manually curated to ensure that the model was able to generate feasible solutions (Sánchez et al., 2017). The second iteration of GECKO (Domenzain et al., 2022) introduced a heuristic for correction of k_{cat} values that is based the on control coefficient of an enzyme (ECC), defined as:

$$C_{ij} = \frac{k_{cat}^{ij} \Delta v_{obj}}{v_{obj} \Delta k_{cat}^{ij}},$$

where v_{obj} is the solution for a given objective function, $\Delta k_{cat}^{ij}$is a perturbation of the k_{cat} values induced by an increase of its initial value by 10-fold, and Δv_{obj} is the change in Δv_{obj} given the change of k_{cat}. The ECCs are then used to rank the enzymes in the pcGEM in decreasing order, with the first enzyme in the list being selected to have its k_{cat} value changed to the maximum k_{cat} value that exists in BRENDA. This operation iterates until the pcGEM can achieve the experimental growth rate provided when reconstructing the pcGEM. This k_{cat} correction heuristic was assessed by reconstructing a new version of the ecYeast7 model using GECKO 2.0 and comparing

it to its first version, which was reconstructed with GECKO 1.0. Simulations performed with the model reconstructed with GECKO 2.0 had a lower average relative error than the model reconstructed with GECKO 1.0 when compared to experimental data in 19 different conditions, these being 23.97% of average relative error and 32.07% of average relative error, respectively.

An algorithm for correcting k_{cat} values has been proposed by (Wendering et al., 2022), named PRESTO (protein-abundance-based correction of turnover numbers), which leverages measurements of protein abundance and exchange fluxes over multiple conditions to correct k_{cat} values. Instead of control coefficients, PRESTO uses a linear optimization approach. A correction factor δ is added to the k_{cat} value of each enzyme i, given that protein abundance values are also available:

$$\sum_{r \in GPR(E_i)} v_r \leq \left(k_{cat,i}^{min} + \delta_i\right) E_{ij}, \quad \forall i \in E_{measured}.$$

The objective of the optimization problem is to minimize a weighted linear combination of the average relative error for predicted specific growth rates and the correction of the initial turnover number integrated in the pcGEM:

$$\min_{v,\delta,\omega} \frac{1}{|C|} \sum_{j \in condition} |\omega_j| + \frac{\lambda}{|E_{measured}|} \sum_{i \in E_{measured}} \delta_i,$$

where ω is the average relative error between experimental and predicted growth rates in an experimental condition C and λ is a fitted parameter that controls the trade-off between the minimization objectives. The problem is then defined as:

$$\min_{v,\delta,\omega} \frac{1}{|C|} \sum_{j \in condition} |\omega_j| + \frac{\lambda}{|E_{measured}|} \sum_{i \in E_{measured}} \delta_i$$

$$\text{s.t. } \mathbf{N v_j} = \mathbf{0}, 1 \leq j \leq |C|$$

$$\sum_{r \in GPR(E_i)} v_r \leq \left(k_{cat,i}^{min} + \delta_i\right) E_{ij}, \forall i \in E_{measured}$$

$$\mathbf{v_j^{min}} \leq \mathbf{v_j} \leq \mathbf{v_j^{max}}$$

$$\delta_i \leq (\varepsilon - 1) \cdot k_{cat,i}^{min}$$

$$k_{cat,i}^{min} + \delta_i \leq k^{max}$$

$$v_{bio,j} \cdot \omega_j \geq \mu_j^{exp} - v_{bio,j}$$

$$v_{bio,j} \cdot \omega_j \geq v_{bio,j} - \mu_j^{exp}$$

$$\omega \leq 0, \delta \geq 0.$$

To validate the approach, the corrected k_{cat} values were integrated in the *S. cerevisiae* GEM Yeast8 and the *E. coli* GEM iML1515. The resulting pcGEMs where then used to simulate phenotypes in three conditions: (i) where only the total protein content of a certain growth condition was available, (ii) where uptake constraints were considered alongside total protein content, and (iii) measured protein abundances were also integrated. The models integrated with PRESTO-corrected k_{cat} values displayed lower relative errors for the three simulation conditions than the models integrated with GECKO-corrected k_{cat} values using the pcGEMs ecYeast8 and eciML1515 generated using GECKO 2.0. These findings highlighted that using physiological data and enzyme abundance can enhance estimations of k_{cat} values. The obtained estimations and corrections of catalytic rates can already greatly enhance the predictive capabilities of pcGEMs. The support of some approaches (e.g., GECKO) for the integration of proteomics data can further improve the predictive capabilities of pcGEMs.

Approaches for prediction of protein abundance

Absolute protein abundance allows for integration of constraints from proteomics data in pcGEMs (Adadi et al., 2012; Sánchez et al., 2017). These data are usually derived from mass spectrometry using methods such as spectral counting or peptide intensity-based quantification (Lindemann et al., 2017). However, given the limitations of these methods, many computational approaches have been developed to predict absolute protein abundance from data that are more facile to gather. Many of these approaches are based on transcriptomics data, sequence-derived data, physicochemical data, or a combination of all. These approaches are usually developed using statistical or supervised learning models, but there are also approaches that stem from the constraint-based modelling framework.

Predictions using data-driven models

One of the earliest attempts at a data-driven model to predict protein abundance was the work of Nie et al. (2006), who developed a zero-inflated Poisson regression model (ZIP) integrating microarray and relative protein abundance data. The data was obtained experimentally by growing *Desulfovibrio vulgaris* on lactate or formate as carbon sources. To build the model, it was assumed that protein abundance *y* follows

a Poisson regression distribution with probability $1 - p$ and a mean λ, and is dependent on mRNA abundance x. The model then is defined as:

$$f(y) = [p + (1 - p) \cdot exp(-\lambda)]^{\delta} \left[(1 - p)exp(-\lambda)\frac{\lambda^y}{y!}\right]^{1-\delta},$$

where δ is 1 if $y = 1$, or 0 if $y \neq 0$. To evaluate the model, they calculated the coefficient of variation (CV) of 30 ribosomal proteins and seven subunits of ATP synthase, and compared it to the CV values of the entire *D. vulgaris* set of proteins.

While the ZIP model could generate predictions with lower CV values for some proteins related to central metabolic pathways or certain operons (Nie et al., 2006), the model relied solely on mRNA and protein data, assuming a linear relationship between the two, which limits its predictive capabilities. Torres-García et al. (2009) proposed instead a non-linear approach, using gradient boosted trees (GBT) and a dataset composed of the microarray and proteomics data from (Nie et al., 2006), numerical sequence-derived data such as protein length, molecular weight, GC content and codon composition, along with categorical data containing the functional category of each protein. Model performance was assessed by the coefficient of determination that ranged from 0.393 to 0.582 and coefficient of variation, which was smaller than that from the model of Nie et al. (2006).

Although the GBT model provided an improvement over the ZIP model, the coefficient of determination was still low. To address this problem, Li et al. (2011) developed a model using a multilayer perceptron (MLP) using the dataset constructed by Torres-García et al. (2009). The resulting MLP is a feed-forward network, using a hyperbolic tangent function as the activating function. It has an input layer that uses the transcriptomics data, a single hidden layer with 6 to 9 neurons (depending on the dataset), and an output layer that would yield the protein abundance values. The coefficient of determination of the trained model ranged from 0.47 to 0.69, representing an improvement over the previous GBT model attempt.

Another data-driven model to predict protein abundances is the Bayesian network constructed by Mehdi et al. (2014) that combined data, consisting of mRNA expression levels, tRNA adaptation index, protein and mRNA half-life, mRNA folding energy, and mRNA interactions with RNA-binding proteins, from *Saccharomyces cerevisiae* and *Schizosaccharomyces pombe*. Model performance was assessed by the Spearman's correlation that ranged from 0.61 to 0.77.

Joint learning approaches have also been used to predict protein abundances. Li et al. (2019) developed an integrated approach to predict protein abundance in breast and ovarian cancer cells. The devised approach contains three parts. First, a generic model learns the relationship between mRNA expression and protein abundance. Then, a series of protein-specific random forest models are used to learn how individual genes behave in a network. Lastly, a cross-sample model uses combined data of the two cancer cells, and is likewise trained as the protein-specific random forest models. An ensemble of the three parts was then trained, using the weighted average predictions from each part to predict protein abundance. This approach outperformed other approaches in the NCI-CPTAC DREAM Proteogenomics Challenge, and the predictions achieved an average Pearson correlation of 0.53.

Besides mRNA expression, mRNA secondary structures have also been used to predict protein abundances. Terai & Asai (2020) developed RBSeval, an approach trained with three different algorithms to predict protein abundance in *Escherichia coli*, using features such as accessibility around the Shine-Dalgarno sequence, minimum free energy of the mRNA molecule, Viterbi score, and inside-outside score, those being calculated either by the Turner model or the CONTRAfold model. The model was assessed by Spearman's correlation, which ranged from 0.554 to 0.709.

The data-driven approaches so far have been heavily reliant on experimental data for training the models, which can become troublesome if these approaches are to be applied to organisms different from those used in the original manuscripts. In this regard, Ferreira et al. (2021) trained an AdaBoost regression model to predict protein abundances using codon usage metrics as features. The model was trained on *Saccharomyces cerevisiae* proteomics data, with predictions achieving a Spearman's correlation of 0.744 when compared to experimental data. The model was then used to predict protein abundance for *E. coli*, *Schizosaccharomyces pombe*, and *Kluyveromyces marxianus*, achieving Spearman's correlations of 0.503, 0.702 and 0.623, respectively. Predictions from *Saccharomyces cerevisiae* where also assessed by integrating the predicted protein abundances in the ecYeast8 pcGEM, which yielded metabolic flux simulations in agreement with simulations performed using experimental proteomics data.

While the approach from Ferreira et al (2021) and from previous models achieved good predictions, they still relied on data from optimal growth conditions. In

addition, given that the proteome is remodelled when physiological and/or environmental changes occur, machine learning models that rely on static features (such codon usage, macromolecular function and structure, housekeeping gene expression) cannot predict the dynamic nature of the proteome. Further, current constraint-based methods underestimate predictions of protein abundance (See Sections 5.2 and 6.1). To address these issues, an integrated framework of constraint-based modelling with machine learning, named CAMEL (Coupled Approach of MEtabolic modelling and machine Learning), was recently developed (Moura Ferreira et al., 2023). The constraint-based module of CAMEL predicts the enzyme usage distribution and the flux distribution under a certain growth condition. The predicted enzyme usage distribution is employed along with experimental proteomics measurements to calculate the protein reserve ratio, which is the discrepancy between measured and predicted protein abundances. The protein reserve ratio is then used to train machine learning models using the TPOT automated tool (Olson and Moore, 2019), using as features the enzyme usage and flux distributions, and codon usage metrics calculated from coding sequences. By employing the predicted protein reserve ratios, it was possible to calculate the *in vivo* protein abundances, matching the experimental proteomics measurements, since both its ratio relative to predictions and the predictions themselves are known. For *E. coli*, the CAMEL-calculated *in vivo* protein abundances achieved a Pearson correlation to experimental proteomics measurements of over 0.9, while for *S. cerevisiae*, it obtained a correlation of 0.5.

The described data-driven models also vary widely in how they are developed, how accessible is the source code, and how accessible is their documentation. The joint learning approach from Li et al. (2019) has its entire code and datasets available on GitHub, with basic instructions to reproduce the analysis of the manuscript. The software tool RBSeval is also available on GitHub, and is implemented as a command-line tool, requiring as input only a FASTA file of coding sequences. However, since it depends on features only available on prokaryotes (e.g., Shine-Dalgarno sequence), it is not applicable with data from eukaryotic organisms. The code and data for the predictive models of Ferreira et al. (2021) are also available on GitHub, including code for reproducing the analysis in its manuscript. The CAMEL approach also has its code and data available on a GitHub repository, for both the constraint-based part and the

machine learning part, allowing for reproduction of the manuscript's findings and for usage for other organisms.

Predictions using constraint-based models

Apart from data-driven models, constraint-based models have also been employed to predict protein abundance. More specifically, they predict enzyme concentration in pcGEMs or resource allocation models, given the relation $v_j \leq k_{cat}^{ij} \cdot [E_i]$ for *in vitro* k_{cat} values, or if k_{app} values are available, $v_j(C) = k_{app,j} \cdot \eta(C)$. Goelzer et al. (2015) reconstructed an RBA model of *Bacillus subtilis* to study its physiological processes. They obtained predictions of fluxes, enzyme abundances, and resource costs of cellular processes. Regarding the predictions of enzyme concentrations, they obtained a R^2 value of 0.94 for growth simulations when using a minimal medium added with pyruvate, glucose, or a combination of glucose and glutamate as carbon sources. The framework developed by Heckmann et al. (2018) was also used for the prediction of protein abundance, given that the machine learning-predicted $k_{app,\max}$ values were used in the pcGEMs. This was used to assess the improvement of using $k_{app,\max}$ values over k_{cat} values, which resulted in a prediction error 43% lower for the former when using MOMENT or ME-models. Similarly, the k_{cat} values predicted using the tools DLKcat (Li et al., 2022) and TurNuP (Kroll et al., 2023) were also used to calculate protein abundances, and both approaches were compared to the same experimental data. For DLKcat, pcGEMs parameterized with its predicted k_{cat} values achieved a root mean squared error 30% lower than pcGEMs parameterized with their original k_{cat} values. For TurNuP, pcGEMs parameterized with its predicted k_{cat} values could predict protein abundance more accurately than DLKcat in 19 out of 21 growth conditions, with an average of 18% lower mean squared errors between measured and predicted protein abundances.

The described constrained-based approaches relied on predicting protein abundances by directly considering the relation between k_{cat} values, protein abundances and metabolic fluxes that reflect a given physiological state. Changes in growth conditions are often accompanied by changes in the allocation of proteins, to facilitate establishment of new homeostasis. Two approaches based on the minimization of metabolic adjustment principles have been developed to specifically

predict the adjustment of enzyme usage. The first approach, termed PARROT (Ferreira et al., 2023), proposes the minimization of the Manhattan (linear) or Euclidean (quadratic) distances between the enzyme usage distribution of a reference growth condition and an alternative growth condition, with or without the consideration of metabolic fluxes. PARROT is available on a GitHub repository, and is implemented in MATLAB, as is therefore compatible with pcGEMs generated with the GECKO Toolbox (Sánchez et al., 2017). When compared to experimental proteomics data, PARROT achieved higher Pearson correlations than other methods, such as pFBA, which is the current standard for pcGEMs (Domenzain et al., 2022). The second approach is part of the OVERLAY tool (Yao et al., 2023), which implements an optimization problem to minimize the Euclidean distance between a transcript abundance distribution obtained from RNA-seq data and the enzyme usage distribution, on the assumption that both distributions are similar and that the predicted enzyme usage distribution maintains metabolic feasibility. This approach is thus dependent on the availability of gene expression data, while PARROT has no requirement of additional data to generate predictions. The predicted enzyme usage distribution from OVERLAY highly matched the RNA-seq data, achieving a R^2 score of over 0.9. However, given that the resulting enzyme usage distribution is simply the result of a distance minimization to the RNA-seq data, this result is an artefact of fitting the RNA-seq data as part of the approach.

While current pcGEMs have expanded our understanding of enzyme usage and allocation, allowing for multi-omics data analysis and design of metabolic engineering strategies, there is still a lot of room for improvement, especially for use cases not assessed by the proposed methods (e.g., metabolic engineering strategies directed at improvement of enzyme catalytic properties, coupled with over/under-expression of enzyme abundances). In addition, some standing questions remain unanswered, which should guide the development of new approaches for estimation and integration of enzyme constraints.

Future directions for estimation and integration of enzyme constraints: standing questions and new opportunities

Do the predicted protein abundances match protein allocation?

Cells produce enzymes in higher quantities than necessary to sustain growth, meaning that they do not operate at their full catalytic capacity (O'Brien et al., 2016). This over-abundance of enzymes could act as a buffer to allow the cell to quickly adapt to variation in nutrient uptakes or other changes in the environment (Mori et al., 2017). However, this also raises the question of whether predictions of protein abundance can match the amount of protein that is actually allocated by the cell. Using the relationship $v_{max} = k_{cat} \cdot [E_i]$ will yield $[E_i]$ values that correspond to the optimal abundance of enzymes necessary to carry the provided flux with the provided catalytic rate. This means that these calculations underestimate *in vivo* enzyme concentrations. As a result, it poses a challenge even for integration of protein abundance values in pcGEMs, since not the entire enzyme pool is used to carry flux, rendering enzymes unsaturated. Thus, a range of enzyme usage values could allow the pcGEM to provide similar predictions, as metabolite levels would determine the alteration of metabolic fluxes. This raises the question of how important absolute protein measurements actually are, since in scenarios in which metabolic fluxes reach respective thresholds, increasing protein abundance would have negligible effect on flux variability. Addressing this question would require directing research efforts towards obtaining proteomics data from scenarios in which enzyme usage thresholds are met.

Along these lines, models that integrate macromolecular machineries and resource allocation strategies, such as ME-models and RBA models, can capture the protein allocation principles, but they depend on biochemical parameters that can be difficult to obtain. Data-driven models could be used to overcome this problem, as features used to train the models such as mRNA expression and codon usage bias can be a proxy for *in vivo* enzyme concentrations. A coupling of data-driven and constraint-based models could further enhance predictions, as proposed in the CAMEL approach (Ferreira et al., 2023), where the ratio between pcGEM-predicted protein abundance values and *in vivo* measurements were used to train machine learning models. This development highlights the opportunities for integrating multi-omics and multi-model approaches.

Towards protein-constrained models of microbial communities

In natural habitats, microbes seldom live isolated. Instead, they engage in complex social interactions that shape their ecosystem (Konopka, 2009). GEMs of microbial

communities have also been reconstructed and analysed using FBA and its derivations, and they provide valuable insights on how these communities are organized and how they function at the metabolic level (Ibrahim et al., 2021). However, integration and estimations of turnover numbers and enzyme abundances have been so far limited to GEMs of single organisms, and absolute quantification of proteins in microbial communities is still an under-explored venue. Possible challenges for reconstructing pcGEMs for microbial communities include reliance on data from single organisms, such as the k_{cat} values deposited on BRENDA or SABIO-RK, and the lack of data for *in silico* estimation of these parameters. Even if data were available, the taxonomic heterogeneity of species in the GEM can make it difficult to map turnover numbers to the reactions, as k_{cat} values can be different from one species to another, which could impact simulations. Community models can be reconstructed by either combining all the overlapping genetic and metabolic information as if it were a single organism, or by reconstructing many small individual models that correspond to one species, where the small individual models are treated as compartments of a bigger model with a shared extracellular compartment (Dillard et al., 2021). Therefore, the structure of the community model could also impact how enzymatic constraints are integrated. A possible workaround for integrating k_{cat} values in community models could be the representation of reactions as devised by Bulović et al. (2019) for RBA models, where they bypass the need of specifying parameters for each protein individually by using the BiPON ontology (Henry et al., 2017) – an annotation that represents cellular processes in a unified way, to represent overlapping processes as one entity, adhering to what is done for community models when reactions shared between species are lumped to a single reaction. This opens the opportunity for novel approaches in integrating enzymatic constraints in community models.

Heterogeneity of cell types in higher eukaryotes

The integration of k_{cat} values for most prokaryotes and single-celled eukaryotes is straightforward, as there are no variations in cell type that could drastically change enzyme activity when calculating k_{cat} values from estimated fluxes (Ohno et al., 2008). In multicellular organisms such as plants and animals, however, cells can differentiate and form tissues and organs, with each cell type finely tuned to a specific metabolism, which could present a challenge for integration of k_{cat} values in GEMs of multicellular

organisms. One workaround to this problem is developing a cell-agnostic model, which is reconstructed using all reactions from all cell types and tissues. This cell-agnostic model can then be constrained using transcriptomics or proteomics to generate tissue-specific models, such as the 126 models of human tissues reconstructed using the mCADRE tool (Wang et al., 2012); the 11 tissue-specific human models and models of guard cells and mesophyll cells of *Arabidopsis thaliana* reconstructed using the RegrEx approach, which also provides flux distributions (Robaina-Estévez et al., 2017; Robaina-Estévez & Nikoloski, 2014). By using protein measurements from multiple tissues, one could obtain estimates of catalytic rates that correspond to a specific cell type, allowing for more accurate reconstructions of tissue-specific pcGEMs. A question that arises is whether such k_{cat} values would be dependent of the tissue context, as different cell types – e.g., photosynthetically-active mesophyll cells vs. starch-accumulating root cells – have evolved cellular phenotypes with different metabolic objectives and may thus possess different metabolic environments where kinetic properties can differ.

Usage of fluxes determined from labelling studies

An interesting point to consider is the usage of fluxes determined from labelling studies to estimate k_{cat} values, rather than using fluxes determined from constraint-based approaches like FBA or pFBA. Experimental flux estimation comes from ^{13}C-based metabolic flux analysis (^{13}C-MFA), which makes use of tracers labelled with ^{13}C, such as a carbon source used for growth (Zamboni et al., 2009). As this tracer gets consumed and integrated into other metabolites, a particular labelling pattern will be achieved. The metabolites that incorporate the tracer can then be detected by analytical techniques such as nuclear magnetic resonance or mass spectrometry (Fischer et al., 2004; Truong et al., 2014). Internal *in vivo* flux estimations can be derived by applying these measurements to a small scale metabolic model and solving a non-linear least squared regression problem (Sokolenko et al., 2016). Using flux values estimated from ^{13}C-MFA comes with advantages such as having more reliable and precise measurements of metabolic fluxes (Crown and Antoniewicz, 2013). Using ^{13}C-MFA data, estimations of k_{cat} should also be of higher accuracy. However, this method comes with challenges such as the resource intensiveness of obtaining the data and the flux estimates (due to the large non-linear optimization problems solved

(Sokolenko et al., 2016)). Experimentally, the analytical steps require advanced training and expensive instrumentation, and the scale of estimations is still limited, being often confined to central metabolic pathways (Ohno et al., 2022). All in all, as ^{13}C-MFA becomes more widely accessible, k_{cat} estimations can take advantage of the more precise flux estimations, leading to even better phenotype predictions by pcGEMs.

Underground metabolism and promiscuous enzymes

Many enzymes display side activities by catalysing reactions other than its main reaction. These enzymes are known as promiscuous, and they play an important role in metabolism (Amin et al., 2019). In the cell context, promiscuous enzymes can form an alternative metabolic network of reactions termed underground metabolism. While many of these reactions are physiologically irrelevant given the low enzymatic activity, they can act as a reservoir of novel enzyme functions that can arise in circumstance where the side reaction is favoured (Rosenberg and Commichau, 2019). Evolution of such novel functions can also be exploited for biotechnological purposes in adaptive laboratory evolution (ALE) experiments (Kovács et al., 2022). In a constraint-based modelling context, promiscuous enzymes affect the GPR rules in a way that multiple reactions will be catalysed by the same enzyme. This makes it challenging to integrate k_{cat} values for promiscuous enzymes, as databases might not contain information about non-canonical enzyme/substrate pairs, since the difference in substrate affinity means each reaction will have different k_{cat} values (Davidi et al., 2016). Experimental estimation of catalytic rates of promiscuous enzymes also run into many roadblocks, because the products of side reactions might be unknown, and the yield of such products be undetectably low (Waki et al., 2021). Even if data would be available, many approaches do not have mechanisms to deal with promiscuous enzymes. Approaches such as MOMENT and PAM assign a single k_{cat} value to the enzyme irrespective of the substrate. The eMOMENT approach explicitly account for enzyme promiscuity by adding a constraint that limits the abundance of an enzyme to the sum of all enzyme abundances across their respective associated reactions, but still doesn't take into account substrate affinity. Nevertheless, data-driven approaches such as DLKcat have shown good promise of estimating catalytic rates of promiscuous enzymes (Li et al., 2022). As biochemical and computational methods become more refined, the inflow of

data might push for the development of novel approaches for considering promiscuous activity in pcGEMs. This could enhance the simulation of ALE experiments and further boost metabolic engineering endeavours.

Conclusion

From 191 catalytic rates estimated by Valgepea et al. (2013) to more than 300.000 catalytic rates estimated by Li et al. (2022), there have been great strides in the parametrization of pcGEMs, with many methods for estimating enzyme kinetics and enzyme abundance being developed and achieving good agreement with experimental measurements. The many approaches developed to integrate catalytic rates and enzyme usage in GEMs have contributed significantly to help understand complex phenotypes and assist in metabolic engineering endeavours. They also provide an opportunity in which multiple omics datasets can be integrated. Some challenges still remain, though, which should provide a fulcrum to future research.

Acknowledgements

This study was financed in part by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior – Brasil (CAPES) – Finance Code 001. We thank Eduardo Almeida, Marius Arend and Philipp Wendering for their critical discussion and comments on this study.

References

Adadi R, Volkmer B, Milo R, Heinemann M, Shlomi T. 2012. Prediction of Microbial Growth Rate versus Biomass Yield by a Metabolic Network with Kinetic Parameters. Ed. Nathan D. Price. *PLoS Comput Biol* **8**:e1002575.

Alter TB, Blank LM, Ebert BE. 2021. Proteome Regulation Patterns Determine Escherichia coli Wild-Type and Mutant Phenotypes. *mSystems* **6**. https://journals.asm.org/doi/10.1128/mSystems.00625-20.

de Alteriis E, Cartenì F, Parascandola P, Serpa J, Mazzoleni S. 2018. Revisiting the Crabtree/Warburg effect in a dynamic perspective: a fitness advantage against sugar-induced cell death. *Cell Cycle* **17**:688–701.

Amin SA, Chavez E, Porokhin V, Nair NU, Hassoun S. 2019. Towards creating an extended metabolic model (EMM) for E. coli using enzyme promiscuity prediction and metabolomics data. *Microb Cell Fact* **18**:1–12.

Basan M, Hui S, Okano H, Zhang Z, Shen Y, Williamson JR, Hwa T. 2015. Overflow metabolism in Escherichia coli results from efficient proteome allocation. *Nature* **528**:99–104.

De Becker K, Totis N, Bernaerts K, Waldherr S. 2022. Using resource constraints derived from genomic and proteomic data in metabolic network models. *Curr Opin Syst Biol*. Elsevier.

Beg QK, Vazquez A, Ernst J, De Menezes MA, Bar-Joseph Z, Barabási AL, Oltvai ZN. 2007. 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.

Bekiaris PS, Klamt S. 2020. Automatic construction of metabolic models with enzyme constraints. *BMC Bioinformatics* **21**:1–13. https://bmcbioinformatics.biomedcentral.com/articles/10.1186/s12859-019-3329-9.

Bulović A, Fischer S, Dinh M, Golib F, Liebermeister W, Poirier C, Tournier L, Klipp E, Fromion V, Goelzer A. 2019. Automated generation of bacterial resource allocation models. *Metab Eng* **55**:12–22.

Calderón-Celis F, Encinar JR, Sanz-Medel A. 2018. Standardization approaches in absolute quantitative proteomics with mass spectrometry. *Mass Spectrom Rev* **37**:715–737.

Chen Y, Nielsen J. 2021. In vitro turnover numbers do not reflect in vivo activities of yeast enzymes. *Proc Natl Acad Sci U S A* **118**:e2108391118. https://www.pnas.org/doi/abs/10.1073/pnas.2108391118.

Crown SB, Antoniewicz MR. 2013. Publishing 13C metabolic flux analysis studies: A review and future perspectives. *Metab Eng* **20**:42–48.

Davidi D, Noor E, Liebermeister W, Bar-Even A, Flamholz A, Tummler K, Barenholz U, Goldenfeld M, Shlomi T, Milo R. 2016. Global characterization of in vivo enzyme catalytic rates and their correspondence to in vitro kcat measurements. *Proc Natl Acad Sci U S A* **113**:3401–3406. https://www.pnas.org/doi/abs/10.1073/pnas.1514240113.

Dillard LR, Payne DD, Papin JA. 2021. Mechanistic models of microbial community metabolism. *Mol Omics* **17**:365–375.

Domenzain I, Sánchez B, Anton M, Kerkhoven EJ, Millán-Oropeza A, Henry C, Siewers V, Morrissey JP, Sonnenschein N, Nielsen J. 2022. Reconstruction of a catalogue of genome-scale metabolic models with enzymatic constraints using GECKO 2.0. *Nat Commun* **13**:1–13. https://www.nature.com/articles/s41467-022-31421-1.

Ebrahim A, Lerman JA, Palsson BO, Hyduke DR. 2013. COBRApy: COnstraints-Based Reconstruction and Analysis for Python. *BMC Syst Biol* **7**:1–6. https://bmcsystbiol.biomedcentral.com/articles/10.1186/1752-0509-7-74.

Ferreira MA de M, Silveira WB da, Nikoloski Z. 2023. PARROT: Prediction of enzyme abundances using protein-constrained metabolic models. Ed. Christos A. Ouzounis. *PLoS Comput Biol* **19**:e1011549. https://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.1011549.

Ferreira M, Ventorim R, Almeida E, Silveira S, Silveira W. 2021. Protein Abundance Prediction Through Machine Learning Methods. *J Mol Biol* **433**:167267.

Fischer E, Zamboni N, Sauer U. 2004. High-throughput metabolic flux analysis based on gas chromatography–mass spectrometry derived 13C constraints. *Anal Biochem* **325**:308–316.

Gianchandani EP, Chavali AK, Papin JA. 2010. The application of flux balance analysis in systems biology. *Wiley Interdiscip Rev Syst Biol Med* **2**:372–382.

Goelzer A, Muntel J, Chubukov V, Jules M, Prestel E, Nölker R, Mariadassou M, Aymerich S, Hecker M, Noirot P, Becher D, Fromion V. 2015. Quantitative prediction of genome-wide resource allocation in bacteria. *Metab Eng* **32**:232–243.

Heck M, Neely BA. 2020. Proteomics in Non-model Organisms: A New Analytical Frontier. *J Proteome Res*. NIH Public Access.

Heckmann D, Campeau A, Lloyd CJ, Phaneuf P V., Hefner Y, Carrillo-Terrazas M, Feist AM, Gonzalez DJ, Palsson BO. 2020. Kinetic profiling of metabolic specialists demonstrates stability and consistency of in vivo enzyme turnover numbers. *Proc Natl Acad Sci U S A* **117**:23182–23190. https://www.pnas.org/doi/abs/10.1073/pnas.2001562117.

Heckmann D, Lloyd CJ, Mih N, Ha Y, Zielinski DC, Haiman ZB, Desouki AA, Lercher MJ, Palsson BO. 2018. Machine learning applied to enzyme turnover numbers

reveals protein structural correlates and improves metabolic models. *Nat Commun* **9**:5252.

Henry VJ, Goelzer A, Ferré A, Fischer S, Dinh M, Loux V, Froidevaux C, Fromion V. 2017. The bacterial interlocked process ONtology (BiPON): A systemic multi-scale unified representation of biological processes in prokaryotes. *J Biomed Semantics* **8**:1–16.

Ibrahim M, Raajaraam L, Raman K. 2021. Modelling microbial communities: Harnessing consortia for biotechnological applications. *Comput Struct Biotechnol J* **19**:3892–3907.

Kerkhoven EJ. 2022. Advances in constraint-based models: methods for improved predictive power based on resource allocation constraints. *Curr Opin Microbiol* **68**:102168.

Konopka A. 2009. What is microbial community ecology? *The ISME Journal 2009 3:11* **3**:1223–1230.

Kovács SC, Szappanos B, Tengölics R, Notebaart RA, Papp B. 2022. Underground metabolism as a rich reservoir for pathway engineering. *Bioinformatics* **38**:3070–3077.

Kroll A, Rousset Y, Hu XP, Liebrand NA, Lercher MJ. 2023. Turnover number predictions for kinetically uncharacterized enzymes using machine and deep learning. *Nature Communications 2023 14:1* **14**:1–14. https://www.nature.com/articles/s41467-023-39840-4.

Lahtvee PJ, Sánchez BJ, Smialowska A, Kasvandik S, Elsemman IE, Gatto F, Nielsen J. 2017. Absolute Quantification of Protein and mRNA Abundances Demonstrate Variability in Gene-Specific Translation Efficiency in Yeast. *Cell Syst* **4**:495-504.e5.

Li F, Chen Y, Anton M, Nielsen J. 2023. GotEnzymes: an extensive database of enzyme parameter predictions. *Nucleic Acids Res* **51**:D583–D586. https://dx.doi.org/10.1093/nar/gkac831.

Li F, Yuan L, Lu H, Li G, Chen Y, Engqvist MKM, Kerkhoven EJ, Nielsen J. 2022a. Deep learning-based kcat prediction enables improved enzyme-constrained model reconstruction. *Nat Catal*:1–11.

Li F, Nie L, Wu G, Qiao J, Zhang W. 2011. Prediction and Characterization of Missing Proteomic Data in *Desulfovibrio vulgaris*. Ed. E Hovig. *Comp Funct Genomics* **2011**:780973.

Li H, Siddiqui O, Zhang H, Guan Y. 2019. Joint learning improves protein abundance prediction in cancers. *BMC Biol* **17**:1–14. https://bmcbiol.biomedcentral.com/articles/10.1186/s12915-019-0730-9.

Li Z, Nees M, Bettenbrock K, Rinas U. 2022b. Is energy excess the initial trigger of carbon overflow metabolism? Transcriptional network response of carbon-limited Escherichia coli to transient carbon excess. *Microb Cell Fact* **21**:1–19.

Lindemann C, Thomanek N, Hundt F, Lerari T, Meyer HE, Wolters D, Marcus K. 2017. Strategies in relative and absolute quantitative mass spectrometry based proteomics. *Biol Chem* **398**:687–699.

Lloyd CJ, Ebrahim A, Yang L, King ZA, Catoiu E, O'Brien EJ, Liu JK, Palsson BO. 2018. COBRAme: A computational framework for genome-scale models of metabolism and gene expression. *PLoS Comput Biol* **14**:e1006302.

Machado D, Herrgård MJ, Rocha I. 2016. Stoichiometric Representation of Gene–Protein–Reaction Associations Leverages Constraint-Based Analysis from Reaction to Gene-Level Phenotype Prediction. *PLoS Comput Biol* **12**:e1005140.

Mahadevan R, Edwards JS, Doyle FJ. 2002. Dynamic flux balance analysis of diauxic growth in Escherichia coli. *Biophys J* **83**:1331. /pmc/articles/PMC1302231/?report=abstract.

Mao Z, Zhao X, Yang X, Zhang P, Du J, Yuan Q, Ma H. 2022. ECMpy, a Simplified Workflow for Constructing Enzymatic Constrained Metabolic Network Model. *Biomolecules* **12**:65.

Mehdi AM, Patrick R, Bailey TL, Boden M. 2014. Predicting the dynamics of protein abundance. *Molecular and Cellular Proteomics* **13**:1330–1340.

Millán-Oropeza A, Blein-Nicolas M, Monnet V, Zivy M, Henry C. 2022. Comparison of Different Label-Free Techniques for the Semi-Absolute Quantification of Protein Abundance. *Proteomes* **10**:2.

Moreno-Paz S, Schmitz J, Martins dos Santos VAP, Suarez-Diez M. 2022. Enzyme-constrained models predict the dynamics of Saccharomyces cerevisiae growth in continuous, batch and fed-batch bioreactors. *Microb Biotechnol* **15**:1434–1445.

Mori M, Hwa T, Martin OC, De Martino A, Marinari E. 2016. Constrained Allocation Flux Balance Analysis. *PLoS Comput Biol* **12**:e1004913.

Mori M, Schink S, Erickson DW, Gerland U, Hwa T. 2017. Quantifying the benefit of a proteome reserve in fluctuating environments. *Nature Communications 2017 8:1* **8**:1–8.

Moura Ferreira MA de, Wendering P, Arend M, Batista da Silveira W, Nikoloski Z. 2023. Accurate prediction of in vivo protein abundances by coupling constraint-based modelling and machine learning. *Metab Eng* **80**:184–192. https://linkinghub.elsevier.com/retrieve/pii/S1096717623001404.

Nie L, Wu G, Brockman FJ, Zhang W. 2006. Integrated analysis of transcriptomic and proteomic data of Desulfovibrio vulgaris: zero-inflated Poisson regression models to predict abundance of undetected proteins. *Bioinformatics* **22**:1641–1647.

Nilsson A, Nielsen J, Palsson BO. 2017. Metabolic Models of Protein Allocation Call for the Kinetome. *Cell Syst* **5**:538–541.

O'Brien EJ, Utrilla J, Palsson BO. 2016. Quantification and Classification of E. coli Proteome Utilization and Unused Protein Costs across Environments. *PLoS Comput Biol* **12**:e1004998.

Ohno H, Naito Y, Nakajima H, Tomita M. 2008. Construction of a Biological Tissue Model Based on a Single-Cell Model: A Computer Simulation of Metabolic Heterogeneity in the Liver Lobule. *Artif Life* **14**:3–28.

Ohno S, Uematsu S, Kuroda S. 2022. Quantitative metabolic fluxes regulated by trans-omic networks. *Biochemical Journal* **479**:787–804.

Olson RS, Moore JH. 2019. TPOT: A Tree-Based Pipeline Optimization Tool for Automating Machine Learning. In: Hutter, F, Kotthoff, L, Vanschoren, J, editors. *Automated Machine Learning: Methods, Systems, Challenges*. Cham: Springer International Publishing, pp. 151–160. https://doi.org/10.1007/978-3-030-05318-5_8.

Orth JD, Thiele I, Palsson BØ. 2010. What is flux balance analysis? *Nat Biotechnol* **28**:245–248.

Otto A, Becher D, Schmidt F. 2014. Quantitative proteomics in the field of microbiology. *Proteomics* **14**:547–565.

Palsson BØ. 2015. Systems Biology: Constraint-based Reconstruction and Analysis 1st ed. Cambridge: Cambridge University Press 531 p.

Pappireddi N, Martin L, Wühr M. 2019. A Review on Quantitative Multiplexed Proteomics. *ChemBioChem* **20**:1210–1224.

Paszke A, Gross S, Massa F, Lerer A, Bradbury J, Chanan G, Killeen T, Lin Z, Gimelshein N, Antiga L, Desmaison A, Köpf A, Yang E, DeVito Z, Raison M, Tejani A, Chilamkurthy S, Steiner B, Fang L, Bai J, Chintala S. 2019. PyTorch: An Imperative Style, High-Performance Deep Learning Library. *Adv Neural Inf Process Syst* **32**. https://arxiv.org/abs/1912.01703v1.

Regueira A, Lema JM, Mauricio-Iglesias M. 2021. Microbial inefficient substrate use through the perspective of resource allocation models. *Curr Opin Biotechnol* **67**:130–140.

Robaina Estévez S, Nikoloski Z. 2014. Generalized framework for context-specific metabolic model extraction methods. *Front Plant Sci* **5**:491.

Robaina-Estévez S, Daloso DM, Zhang Y, Fernie AR, Nikoloski Z. 2017. Resolving the central metabolism of Arabidopsis guard cells. *Sci Rep* **7**:1–13.

Rosenberg J, Commichau FM. 2019. Harnessing Underground Metabolism for Pathway Development. *Trends Biotechnol*.

Sánchez BJ, Zhang C, Nilsson A, Lahtvee P, Kerkhoven EJ, Nielsen J. 2017. Improving the phenotype predictions of a yeast genome-scale metabolic model by incorporating enzymatic constraints. *Mol Syst Biol* **13**:935.

Shlomi T, Benyamini T, Gottlieb E, Sharan R, Ruppin E. 2011. Genome-Scale Metabolic Modeling Elucidates the Role of Proliferative Adaptation in Causing the Warburg Effect. *PLoS Comput Biol* **7**:e1002018.

Sokolenko S, Quattrociocchi M, Aucoin MG. 2016. Identifying model error in metabolic flux analysis - a generalized least squares approach. *BMC Syst Biol* **10**:1–14.

Swiatly A, Plewa S, Matysiak J, Kokot ZJ. 2018. Mass spectrometry-based proteomics techniques and their application in ovarian cancer research. *J Ovarian Res* **11**:1–13.

Terai G, Asai K. 2020. Improving the prediction accuracy of protein abundance in Escherichia coli using mRNA accessibility. *Nucleic Acids Res* **48**.

Thiele I, Palsson B. 2010. A protocol for generating a high-quality genome-scale metabolic reconstruction. *Nat Protoc* **5**:93–121.

Torres-García W, Zhang W, Runger GC, Johnson RH, Meldrum DR. 2009. Integrative analysis of transcriptomic and proteomic data of Desulfovibrio vulgaris: a non-linear model to predict abundance of undetected proteins. *Bioinformatics* **25**:1905–1914.

Truong QX, Yoon JM, Shanks J V. 2014. Isotopomer measurement techniques in metabolic flux analysis I: Nuclear magnetic resonance. *Methods in Molecular Biology* **1083**:65–83.

Valgepea K, Adamberg K, Seiman A, Vilu R. 2013. Escherichia coli achieves faster growth by increasing catalytic and translation rates of proteins. *Mol Biosyst* **9**:2344–2358. https://pubs.rsc.org/en/content/articlehtml/2013/mb/c3mb70119k.

Valgepea K, Adamberg K, Vilu R. 2011. Decrease of energy spilling in Escherichia coli continuous cultures with rising specific growth rate and carbon wasting. *BMC Syst Biol* **5**:1–11.

Waki T, Takahashi S, Nakayama T. 2021. Managing enzyme promiscuity in plant specialized metabolism: A lesson from flavonoid biosynthesis. *BioEssays* **43**:2000164.

Wang Y, Eddy JA, Price ND. 2012. Reconstruction of genome-scale metabolic models for 126 human tissues using mCADRE. *BMC Syst Biol* **6**:1–16.

Wendering P, Arend M, Razaghi-Moghadamkashani Z, Nikoloski Z. 2022. Data integration across conditions improves turnover number estimates and metabolic predictions. *bioRxiv*:2022.04.01.486742.

Wendering P, Nikoloski Z. 2022. Genome-Scale Modeling Specifies the Metabolic Capabilities of Rhizophagus irregularis. *mSystems* **7**.

Wilken SE, Besançon M, Kratochvíl M, Foko Kuate CA, Trefois C, Gu W, Ebenhöh O. 2022. Interrogating the effect of enzyme kinetics on metabolism using differentiable constraint-based models. *Metab Eng* **74**:72–82.

Xu R, Razaghi-Moghadam Z, Nikoloski Z. 2021. Maximization of non-idle enzymes improves the coverage of the estimated maximal in vivo enzyme catalytic rates in Escherichia coli. *Bioinformatics* **37**:3848–3855. https://academic.oup.com/bioinformatics/article/37/21/3848/6343444.

Yao H, Dahal S, Yang L. 2023. Novel context-specific genome-scale modelling explores the potential of triacylglycerol production by Chlamydomonas reinhardtii. *Microb Cell Fact* **22**:1–16. https://microbialcellfactories.biomedcentral.com/articles/10.1186/s12934-022-02004-y.

Ye C, Wei X, Shi T, Sun X, Xu N, Gao C, Zou W. 2022. Genome-scale metabolic network models: from first-generation to next-generation. *Appl Microbiol Biotechnol* **106**:4907–4920.

Yizhak K, Benyamini T, Liebermeister W, Ruppin E, Shlomi T. 2010. Integrating quantitative proteomics and metabolomics with a genome-scale metabolic network model. *Bioinformatics* **26**:i255–i260.

Yu H, Deng H, He J, Keasling J, Luo X. 2023. Highly accurate enzyme turnover number prediction and enzyme engineering with PreKcat. https://www.researchsquare.com.

Zamboni N, Fendt SM, Rühl M, Sauer U. 2009. 13C-based metabolic flux analysis. *Nature Protocols 2009 4:6* **4**:878–892.

3.2 PARROT: Prediction of enzyme abundances using protein-constrained metabolic models

Text adapted from the unmarked revised manuscript as accepted by the PLoS Computational Biology journal, available at: https://doi.org/10.1371/journal.pcbi.1011549.

Abstract

Protein allocation determines the activity of cellular pathways and affects growth across all organisms. Therefore, different experimental and machine learning approaches have been developed to quantify and predict protein abundance and how they are allocated to different cellular functions, respectively. Yet, despite advances in protein quantification, it remains challenging to predict condition-specific allocation of enzymes in metabolic networks. Here, using protein-constrained metabolic models, we propose a family of constrained-based approaches, termed PARROT, to predict how much of each enzyme is used based on the principle of minimizing the difference between a reference and an alternative growth condition. To this end, PARROT variants model the minimization of enzyme reallocation using four different (combinations of) distance functions. We demonstrate that the PARROT variant that minimizes the Manhattan distance between the enzyme allocation of a reference and an alternative condition outperforms existing approaches based on the parsimonious distribution of fluxes or enzymes for both *Escherichia coli* and *Saccharomyces cerevisiae*. Further, we show that the combined minimization of flux and enzyme allocation adjustment leads to inconsistent predictions. Together, our findings indicate that minimization of protein allocation rather than flux redistribution is a governing principle determining steady-state pathway activity for microorganism grown in alternative growth conditions.

Author summary

Protein allocation determines the activity of cells and affects diverse traits across all organisms. However, prediction of protein allocation, particularly for conditions that do not result at optimal growth and physiology, remains a very challenging problem. In this study, we present an approach called PARROT to predict how cells allocate their proteins in different conditions. We tested different variants of PARROT by considering

different objectives within a constraint-based formulation and by how much resource allocation information is used to guide predictions. We found that minimizing adjustments in protein allocation, rather than flux phenotypes, is a key principle that microorganisms use under alternative growth conditions. By integrating this principle into our approaches and leveraging quantitative proteomics data, PARROT provides more accurate predictions of protein allocation in unseen conditions in comparison to existing contenders. Therefore, PARROT can help in advancing our understanding of protein allocation under different conditions and its physiological implications. Further, we can gain valuable insights into cellular responses and adaptive strategies across different environments.

Introduction

Constraint-based approaches have been employed to simulate and predict phenotypes based on genome-scale metabolic models (GEMs) [1]. While already useful for predicting a wide range of phenotypes, the predictive performance of GEMs has been further improved by integrating protein constraints, such as: enzyme catalytic rates and the allocation of enzyme abundances across reactions [2,3]. Enzyme abundances are central to metabolic function, since they impact the rate of reactions and regulation of pathways, with implications to biotechnology and medicine[4]. These protein-constrained GEMs (pcGEMs) have been used to predict complex phenotypes, such as the overflow metabolism, in which fermentation predominates over respiration when microorganisms grow in high sugar concentrations [3,5], and diauxic growth, when multiple carbon sources are available and the microbial growth presents two or more growth phases [6]. The models also allow for the incorporation of proteomics data, and thus provide a framework for multi-omics data analysis and integration [3,7].

The parameters included in pcGEMs are: (i) the enzyme turnover numbers, k_{cat}, a first-order rate constant with the unit of s^{-1}, that describes the limiting rate of reactions catalysed by enzymes when these are fully occupied at their saturation point; and (ii) enzyme abundances (in mmol/gDW), obtained from quantitative proteomics experiments. Values of k_{cat} can be measured from biochemical assays or estimated from computational methods based on constraint-based and data-driven approaches [8], while enzyme abundances are obtained from absolute proteomics measurements. More specifically, they are obtained from peptide intensity-based quantification or

spectral counting [9]. However, proteomics experiments for absolute quantification are still difficult to perform, given the challenges put forward by the diversity of physico-chemical properties of protein [10], lack of standards and problems in reproducibility [11], and overall inaccessibility given the high costs of equipment and supplies [12].

Computational methods have also been developed to predict protein abundance, mostly based on data-driven models. These models often explore the central dogma of molecular biology by assessing the relationship between transcription and protein biosynthesis. Notable approaches to estimate protein abundance include the joint learning approach devised by Li et al [13], where an ensemble model was constructed by combining different supervised learning algorithms, outperforming competing approaches in the NCI-CPTAC DREAM Proteogenomics Challenge. Another approach, developed by Terai and Asai [14], uses features such as the accessibility around the Shine-Dalgarno sequence, minimum free energy of the mRNA molecule, Viterbi score, and inside-outside score. Further, Ferreira et al. [15] explored codon usage bias information to train an AdaBoost regression model, achieving higher correlations than previous approaches without the usage of transcriptomics data.

Aside from machine learning models, constraint-based approaches have also been used to predict protein abundance. Using approaches such as MOMENT [2] or GECKO [3], it is possible to calculate the optimal concentration of enzymes necessary to carry the provided flux with the provided catalytic rate, given the relationship:

$$v_j \leq k_{cat}^{ij} \cdot [E_i] \qquad (1)$$

where v_j is the metabolic flux of reaction j, $[E_i]$ is the concentration of an enzyme i, and k_{cat}^{ij} is the catalytic rate of an enzyme i catalyzing a reaction j. This allows for deriving k_{cat}^{ij} values given the other two are available. This relationship was explored by Heckmann et al. [16] by using pcGEMs to predict enzyme concentrations given catalytic rates predicted computationally, achieving a 43% lower root mean squared error.

Assuming that pcGEMs that integrate proteomics data predict flux distributions that reflect the corresponding metabolic state, we ask whether the reverse operation could be employed to predict proteomics data that match a given physiological state. Moreover, as cells are exposed to stresses or changing environmental conditions, the current growth state is disturbed, leading to an alternative growth state in which gene

expression, regulatory pathways and metabolic flux are changed in adjusting the cell to the new physiological condition [17]. Despite the aforementioned advances in predicting protein abundances, the problem of predicting enzyme allocation under alternative growth conditions remains largely unexplored. This could be useful to explore how cells adapt to changing environmental conditions, such as those faced by yeasts in industrial fermentations or by pathogenic bacteria when exposed to antibiotics.

Here we propose PARROT (Figure 1), for **P**rotein allocation **A**djustment fo**R** alte**R**native envir**O**nmen**T**s, a family of constraint-based approaches for prediction of protein abundances for alternative growth conditions using protein abundances measured in a reference state. Our proposed approach is inspired by Minimization of Metabolic Adjustment (MOMA) [18], which minimizes the distance between a reference state and a gene knock-out state while ensuring cell survival in the later. We show that PARROT predicted enzyme concentrations in very good agreement with experimental data and outperformed competing methods for minimizing flux distributions. Therefore, PARROT can be used to parameterize pcGEMs for unseen, alternative growth conditions from which metabolic phenotypes can further be analysed.

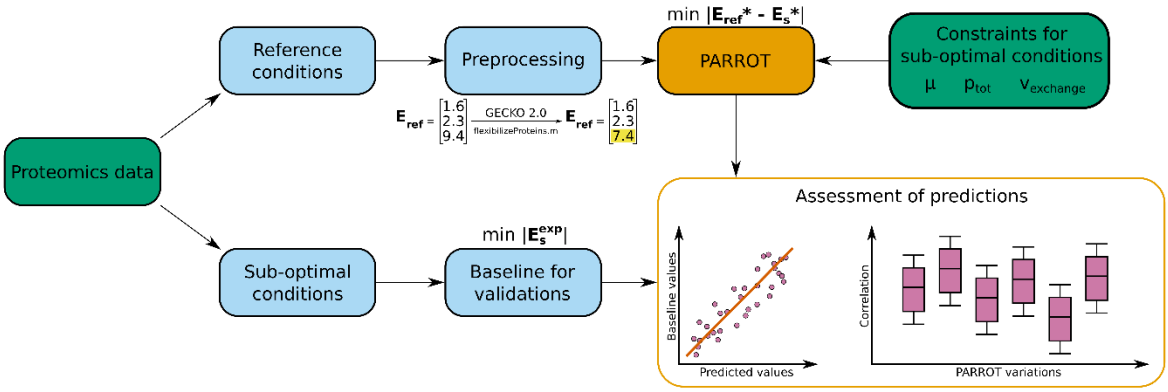

Figure 1. Workflow of PARROT to predict enzyme usage for alternative growth conditions.

PARROT uses experimental proteomics data from a reference growth condition, and experimental physiological parameters from an alternative growth condition in a protein-constrained model. The proteomics data from the reference state is pre-processed by integrating the data in a pcGEM using the GECKO Toolbox 2 and allowing flexibility in its values. The proteomics data from the alternative state is used to generate a baseline, which is in turn used for comparison with predictions from the PARROT variants.

Results

PARROT successfully captures protein allocation changes in yeast

We used PARROT to predict the enzyme usage distribution for 19 growth conditions under constraints provided by experimental data. First, we built a baseline for comparison with predictions from PARROT (Figure 1). To this end, we integrated the experimental proteomics measurements obtained from Lahtvee et al. [19], Yu et al. [20], Di Bartolomeo et al. [21], and Yu et al. [22] (Table S1) in the ecYeast8 model and minimized the enzyme allocation (Methods). The resulting allocation of enzymes $\mathbf{E_s^{exp}}$ included 286 to 336 enzymes with abundance in all considered conditions. For the reference condition, we used the experimental proteomics measurements from the control growth conditions in the respective four groups of experiments, after flexibilization following GECKO 2.0 (see Methods) (Table S1). The number of enzymes contained in $\mathbf{E_{ref}}$ ranged from 533 to 744, depending on the investigated control sample. With the resulting enzyme allocation at the reference and the baseline of an alternative growth condition, $\mathbf{E_{ref}}$ and $\mathbf{E_s^{exp}}$, we used four variants of PARROT (see Methods) to predict the enzyme allocation, $\mathbf{E_s}$, for the alternative growth condition.

The first variant of PARROT (referred as LP1) minimizes the Manhattan distance between $\mathbf{E_{ref}}$ and $\mathbf{E_s}$ (see Methods). The number of enzymes contained in the predicted $\mathbf{E_s}$ ranged from 224 to 253 over the considered experiments. When comparing the median of the calculated Pearson correlations between the baseline and predicted enzyme allocation correlations, we found that LP1 achieved a higher median correlation compared to pFBA and its modified implementation (Figure 2a). This variant also outperformed the null model, where k_{cat} values are used directly as the enzyme usage ($E_{s,i} = k_{cat}^{ij}$). This negative control is useful to determine the contribution of k_{cat} values to the correlation between $\mathbf{E_{exp}}$ and $\mathbf{E_s}$ (see Methods).

The second variant of PARROT, referred as QP1, minimizes the Euclidean distance between $\mathbf{E_{ref}}$ and $\mathbf{E_s}$ (see Methods). The predicted $\mathbf{E_s}$ ranged from 210 to 258 predicted enzymes over the considered experiments. We found that QP1 achieved a higher median correlation when compared to pFBA and its modified implementation, when considering the median of the calculated Pearson correlations between the baseline and $\mathbf{E_s}$ (Figure 2a). As with LP1, this variant outperformed the null model.

The third variant of PARROT, referred as LP2, minimizes the weighted sum of the Manhattan distance between enzyme usage distributions and the Manhattan distance between flux distributions. Thus, this variant also considers the metabolic fluxes of each condition along with the enzyme usage distribution. As observed for the other variants, LP2 outperformed pFBA and its modified implementation, when comparing the median of the calculated Pearson correlations between the baseline and $\mathbf{E_s}$, while also outperforming the null model.

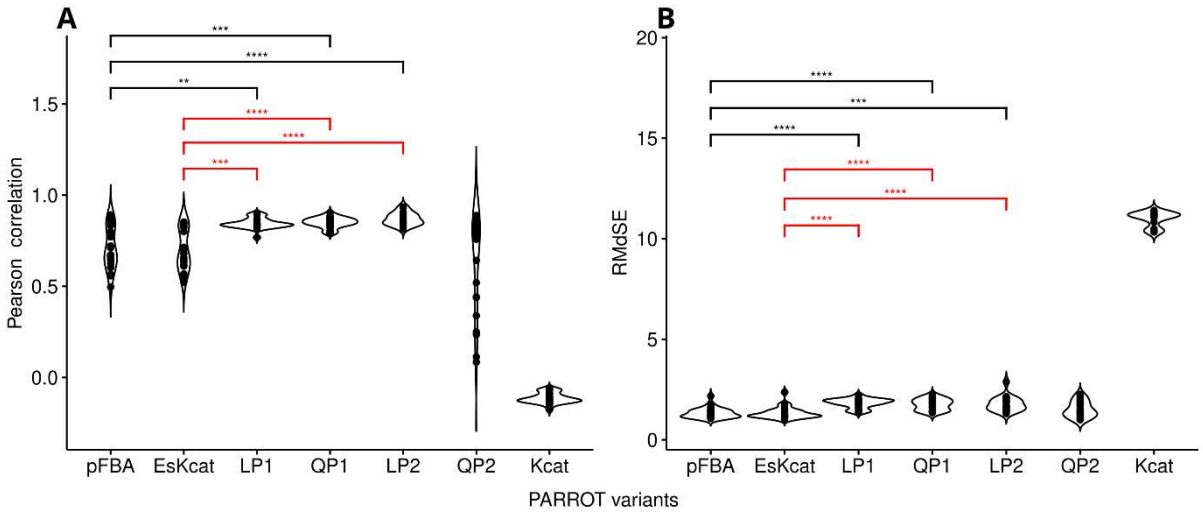

Figure 2. Comparative performance analysis of PARROT with proteomics data from *S. cerevisiae*.

All protein abundance values were log10-transformed prior to comparisons. a. Pearson correlation calculated between predicted enzyme distribution and the baseline obtained from minimizing the first norm of the experimental enzyme usage distribution. The four variants of PARROT are denoted as LP1 (Manhattan distance of enzyme distributions), LP2 (weighted Manhattan distance, considering flux and enzyme distributions), QP1 (Euclidean distance of enzyme distributions), and QP2 (weighted Euclidean distance of flux and enzyme distributions). The performance of PARROT was compared to pFBA and its modified version EsKcat (first norm of enzyme usage), see Methods. A pairwise Wilcoxon rank sum assesses the statistical significance: **** p-value < 1·10-5, *** p-value < 2·10-4, ** p-value < 5·10-4. b. Assessment of model performance based on the root median squared error (RMdSE). A pairwise Wilcoxon rank sum assesses the statistical significance: **** p-value < 9·10-6, *** p-value < 2·10-5. Black significance bar indicates comparisons to pFBA. Red significance bar indicates comparison to EsKcat.

The fourth and final variant of PARROT, QP2, minimizes the weighted sum of the Euclidean distance between enzyme usage distributions and the Euclidean distance between flux distributions. Unlike other variants, QP2 did not achieve a higher median Pearson correlation when comparing the predictions to the baseline, but it was

better than the null model. However, the root median squared error (RMdSE) between predictions and the baseline was the lowest among variants, being comparable to pFBA and its modified implementation (Figure 2b). Taken together, the results demonstrated that PARROT achieved good predictive performance based on the data from *S. cerevisiae* when compared to pFBA and its modified implementation.

Different variants of PARROT outperformed the benchmarks for E. coli

To verify if the conclusions from PARROT hold in another unicellular model organism, we applied it to predict enzyme allocation $\mathbf{E_s}$ in alternative growth conditions for *E. coli* given constraints provided by growth experiments. As in the case of *S. cerevisiae*, we built a baseline for comparison with the predictions obtained from PARROT by integrating the experimental proteomics measurements from Valgepea et al. [23], Peeno et al. [24] and Schmidt et al. [25] (Table S2) in the eciML1515 model, and minimized the total enzyme allocation (see Methods). The resulting $\mathbf{E_s^{exp}}$ included protein allocation for 164 to 176 enzymes. Further, as reference condition we considered the control samples or the chemostat measurements with the smallest dilution rate (Table S2). We chose the smallest dilution rate to ensure that cells are growing aerobically and to prevent the metabolic shifts seen in higher dilution rates (e.g., overflow metabolism). The number of enzymes contained in $\mathbf{E_{ref}}$ ranged from 152 to 188 depending on the control experimented used.

The prediction of $\mathbf{E_s}$ distributions and their assessment were similar to *S. cerevisiae*. The LP1 variant predicted between 122 and 133 enzymes for $\mathbf{E_s}$ across conditions. This variant exhibited significantly higher median correlations compared to pFBA (p-value = $1.24 \cdot 10^{-13}$ for Pearson correlations, pairwise Wilcoxon rank sum test) (Figure 3a). For the QP1 variant, the number of predicted enzymes ranged from 115 to 133 across conditions. Further, it had higher median correlations compared to pFBA, but lower than its modified implementation. The LP2 variant performed similar to LP1, predicting between 125 and 137 enzymes and achieving higher median correlations compared to pFBA and its modified implementation. The variant QP2, on the other hand, predicted a wider range and had a low number of enzymes, ranging from 19 to 141. This variant also had lower median correlations compared to pFBA and its modified implementation. As in *S. cerevisiae*, the QP2 variant had lower RMdSE errors than the other variants. All variants outperformed the null model in all comparisons. These

findings demonstrated that PARROT is applicable with data from another microorganism without decrease in performance.

Robustness analysis shows the consistency of prediction from PARROT

To further evaluate the predictions made by PARROT, we investigated if the construction of the baseline could impact the correlations. Thus, we reconstructed the baseline by minimizing the 2-norm of the vector $\mathbf{E_s^{exp}}$ instead of the 1-norm. To this end, we repeated all comparisons as performed for a baseline constructed by minimizing the 1-norm, using the predicted $\mathbf{E_s}$ obtained by the PARROT variants. Importantly, the results were consistent between the two baseline approaches. For *S. cerevisiae*, the LP2 variant achieved the highest mean Pearson correlations compared to pFBA and its modified implementation (Figure S1). For the RMdSE, all PARROT variants had errors comparable to the positive controls (Figure S2). As observed in the comparisons with the 1-norm baseline, all PARROT variants outperformed the null model.

The comparisons performed using predictions obtained for *E. coli* were also consistent with different variants of PARROT that outperformed pFBA. Considering the Pearson correlations, the LP1 and the LP2 variants also had the highest median correlations and were significantly different to pFBA. Likewise, these PARROT variants also showed significant difference to the modified implementation of pFBA (Figure S3). The comparison of RMdSE values were also consistent with this observation, as the errors were comparable to the positive controls (Figure S4). Altogether, these results highlight the robustness of estimations of $\mathbf{E_s}$ obtained from PARROT.

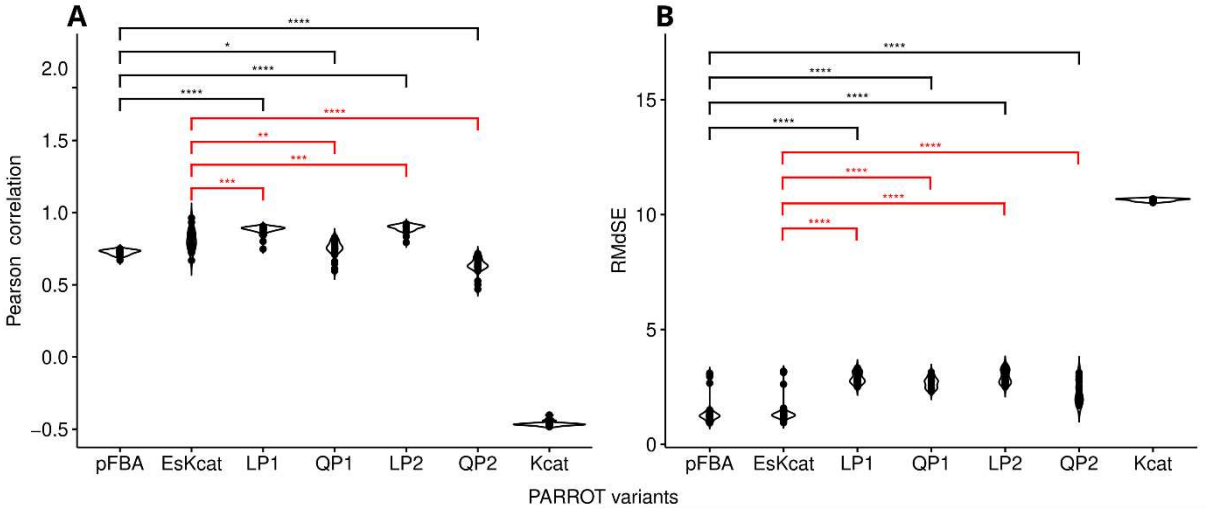

Figure 3. Comparative performance analysis of PARROT with proteomics data from *E. coli.* All protein abundance values were log10-transformed prior to comparisons. **a**. Pearson correlation calculated between predicted enzyme usage distribution and the baseline obtained from minimizing the first norm of the experimental enzyme usage distribution. A pairwise Wilcoxon rank sum assesses the statistical significance: **** p-value < $2 \cdot 10^{-11}$, *** p-value < $2 \cdot 10^{-4}$, ** p-value < $6 \cdot 10^{-3}$, * p-value < $3 \cdot 10^{-2}$. **b**. Assessment of model performance based on the RMdSE in *E. coli*. A pairwise Wilcoxon rank sum assesses the statistical significance: **** p-value < $1 \cdot 10^{-5}$. Black significance bar indicates comparisons to pFBA. Red significance bar indicates comparison to EsKcat.

Proteome-aware minimalization is more relevant than minimization of flux distances

Given that LP2 and QP2 make use of a weighting factor λ, we were interested in how different λ values impact the predictions. We used λ values ranging from 0 (no fluxes used) to 1 (fluxes and enzyme usages equally considered). We also considered a scenario of λ values ranging from 0.1 to 1 in order to probe different solutions where metabolic fluxes are always considered. We considered a λ value to be optimal if it resulted in the highest Pearson correlation to the baseline. In the first scenario, for both *S. cerevisiae* and *E. coli* the most frequent optimal λ was 0, with decreasing correlation values as λ values increased (Figure 4a). In the second scenario, the optimal λ values were more equally distributed, with *S. cerevisiae* having a higher frequency of lower values (Figure 4b). For *E. coli*, lower λ values were also frequent, while also having a λ of 1 slighly more frequent than a λ of 0.2 (Figure 4b). Taken together, these results indicate that the problem of minimizing enzyme usage contributes more to predictions than minimizing metabolic fluxes.

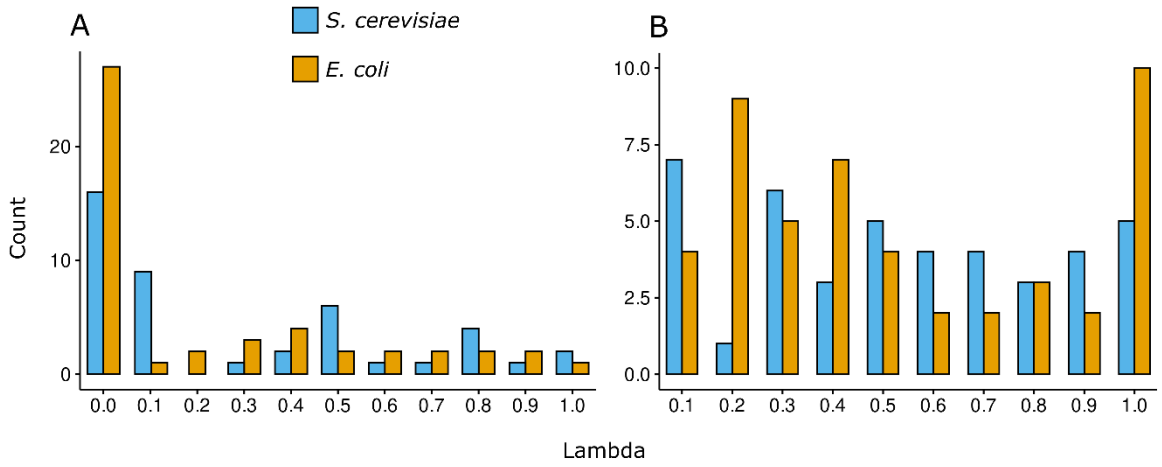

Figure 4. Optimal λ values across conditions and PARROT variants.

The optima λ value was determined by optimising the LP2 and QP2 variants and finding the value that outputs predictions with the highest Pearson correlation when compared to the baseline. Blue bars correspond to *S. cerevisiae*, and orange bars correspond to *E. coli*. a. Number of occurrences of an optimal λ value in a range of 0 to 1. Note that a λ value of zero means that no fluxes are used for the objective, being equivalent to the LP1 and LP2 variants. b. Number of occurrences of an optimal λ value in a range of 0.1 to 1. In this scenario, fluxes are always used for the objective.

Discussion

Here we proposed a family of constraint-based approaches, termed PARROT, that address the problem of predicting reallocation of protein abundance from a reference growth condition to an alternative growth condition. PARROT is based on the principle that organisms tend to minimally adjust cellular physiology between growth conditions to make effective use of resources [26]. The predictions of enzyme allocation generated by PARROT rely on quantitative proteomics data for a reference condition. The resulting optimization problems constructed are thus similar to MOMA, which depends on a model representing a wild-type strain to predict a minimally adjusted flux distribution for a mutant strain.

Understanding how cells adjust enzyme allocation during growth conditions apart from the physiological optimum might prove useful to study, for example, the adaptability of yeasts when exposed to ethanol during fermentation. Ethanol hinders growth and enacts several changes to membrane structure and function, causes protein denaturing and metabolic imbalances [27]. The yeast *Kluyveromyces marxianus*,

important for fermentation of dairy products, adapts to ethanol stress by strengthening the cell membrane by accumulating trehalose and by altering the content of ergosterol and unsaturated fatty acids in the cell membrane, along with changes in several regulatory pathways [28,29]. Another example is the resistance to antibiotics observed in human and animal bacterial pathogens. The prolonged and continuous exposure to antibiotics leads to a selective pressure where the bacterial population acquires resistance to the antibiotic which it is being exposed to, by means of mutations or through acquisition of mobile genetic elements, leading to the emergence of antibiotic-resistant strains [30]. For instance, antibiotic-resistant strains of *Klebsiella pneumoniae*, an ESKAPE pathogen, harbours many plasmids containing genes responsible for coding enzymes that break the antibiotic molecule. An example is the ampC gene family, which codes for β-lactamases, capable of degrading antibiotics such as penicillin, monobactams, cephalosporins and carbapenems. Resistance to these antibiotics is achieved by overexpression of these genes and overproduction of β-lactamases, along with the production of binding proteins that target the antibiotic molecules [31]. Overproduction of enzymes leads to disrupted metabolic states due to the inefficient allocation of resources [32], which can be exploited for therapeutic efforts such as enhancing antibiotic sensitivity [33].

By comparing the predictions to a baseline constructed with experimental proteomics measurements for alternative growth conditions, we found that PARROT predicted protein abundances with very good agreement with the baseline. In addition, we demonstrated that these predictions were consistent and robust to how the baseline is constructed. The performance of PARROT also holds for two model organisms, *S. cerevisiae* and *E. coli*, highlighting the general application of the principle of minimal protein adjustment on which the predictions are based.

From the different variants of PARROT, LP1 (minimization of the Manhattan distance of enzyme usage distributions) and LP2 (the minimization of the weighted sum of the Manhattan distance of enzyme usage and Manhattan distance of flux distributions) were the best contenders across conditions for both *S. cerevisiae* and *E. coli*. The variant QP1 (minimization of the Euclidean distance of enzyme usage distributions) resulted in good, but inconsistent performance between *S. cerevisiae* and *E. coli*. For QP2 (the minimization of the weighted sum of the Euclidean distance of enzyme usage and Euclidean distance of flux distributions), it had poor results for *S.*

cerevisiae, while having good results for *E. coli*, albeit worse than the other variants. This agrees with the fact that the first norm distance is the natural metric for enzyme abundances in the cell, because a change in enzyme concentration requires ribosomal activity that scales linearly with the enzyme abundance [34].

The baseline approach devised to assess the predictions allows for a fair comparison between the predicted enzyme usage distribution and the experimental protein abundance values. In constraining the pcGEMs with the proteomics measurements, the experimental values are first readjusted to match the enzyme levels that actually carry flux in the model, since more protein is produced than actually needed by the cell [35]. This, however, implies that the predicted values are not directly comparable to experimental proteomics values, which affect the determined measures of performance. By adjusting the experimental values to levels that are compatible with what is actually employed to carry metabolic flux, we could more adequately assess the correlation with enzyme allocation predicted from the pcGEMs, albeit losing the direct correspondence to experimental data.

The parameter λ is a factor that weights the usage of metabolic fluxes for the optimisation problem. By varying this value between 0 and 1, we could assess how much the minimization of metabolic fluxes contributes to the problem of predicting enzyme usage. A λ value of 0 would render the variants LP2 and QP2 equivalent to LP1 and QP1, respectively, as metabolic flux would be neglected in the optimal solutions. A λ value of 1, in the other hand, renders LP2 and QP2 as equivalent to using a pcGEM with the canonical implementation of MOMA, which considers all fluxes equally. When the two PARROT variants are free to vary λ between 0 and 1, the optimum is reached for lower λ values. This can be explained by the experimental observation that in changing environments, cells adopt a strategy of initially adjusting gene expression, which subsequently results in shifts in protein allocation. Consequently, this leads to subsequent changes in metabolic flux [21,36]. When constraining λ to a value between 0.1 and 1, higher values of λ are present but still not more prevalent than lower values of λ. This suggests that the joint minimization of fluxes and enzymes is not a principle of flux redistribution, given that when higher λ values are employed in the objective function of LP2 and QP2, the model simultaneously optimizes both the enzyme usage distribution and the flux distribution, as though the cell performs these processes at the same time. Instead, the principle is guided by minimization of resource redistribution,

as best captured by LP1 and QP1, and by LP2 and QP2 with low values of λ, from which a flux redistribution is then later derived. Using lower λ values, the model prioritizes minimizing the enzyme usage distribution, aligning more closely with experimental observations. With a λ of 0, the model disregards metabolic fluxes entirely, enabling it to focus on solving for minimal enzyme usage redistribution, only calculating the flux distribution as a function of the former, mirroring what is observed in the cell [37]. Thus, by being proteome-aware, PARROT is better suited for simulations using pcGEMs than the quadratic and linear implementations of MOMA, given that higher participation of metabolic fluxes lowers the overall predictive performance. Altogether, we demonstrated that minimizing the readjustment of enzyme resource allocation is one principle underpinning microbial adjustment to an alternative condition, aligning with experimental evidence. Thus, PARROT may allow for study and engineering of microbial cell factories, as these are often under suboptimal growth conditions in industrial settings [38].

Despite the advantages of using a baseline, predictions of enzyme levels using Eq. (1) still underestimates protein abundance, leading to a disparity between predictions and *in vivo* concentrations. This remaining portion of proteins, termed the “proteome reserve”, is useful for the cell to quickly adapt to unstable environments, being an evolutionary conserved strategy [39]. It is important to highlight, though, that this reasoning does not assume that cells are operating at the saturation point for all metabolites, but rather that enzymes are used inefficiently. If enzymes are operating near V_{max}, then enzymes would be the only cellular components that exert control on metabolic fluxes. As noted by [40], however, cell overexpress enzymes and uses metabolite concentrations to control metabolic flux. This falls in line with the evolutionary conservation of protein stoichiometries at the pathway level as demonstrated by Lalanne et al. [39]. Although it is still not understood how preferred enzyme stoichiometry is determined, it was observed that the preferred range of enzyme stoichiometry follows a narrow distribution among pathways in Gram-positive and -negative bacteria, likely a result of evolutionary conservation or convergence. As suggested in the study, protein biosynthesis and consequently its usage is bound to a cost-benefit trade-off, where the optimal level of enzymes is balanced with the need for a buffer zone in case of changing environments. Similar to our approach, the works of Mori et al. [37] and Lalanne et al. [39] deals with proteome reallocation in an alternative growth condition. However, the

first deals with proteome sectors, while the latter concerns with pathway-centric stoichiometries. Our approach thus differs as we consider protein reallocation for each enzyme individually.

Nevertheless, other approaches for estimating *in vivo* protein concentrations would still need to overcome the underestimation of protein concentrations of pcGEMs, especially by considering the proteome reserve. Interestingly, Alter et al. [42] deal with the problem of catalytically inactive enzymes by considering a protein sector of unused enzymes, along with active enzymes sector and the translational protein sectors. The unused protein sector, ϕ_{UE}, is calculated by relating the decrease of the concentration of unused enzymes to the increase in the substrate uptake rate. This allows for the model to predict the adaptability of *E. coli* to changing environmental conditions. However, the initial value $\phi_{UE,0}$, which is the unused enzyme concentration when the substrate uptake is zero, is obtained from simulations using a ME model, which are notoriously difficult to parameterize [43]. New approaches could include features such as cellular machinery beyond enzymes that participate in metabolism, or by integrating constraint-based approaches with data-driven approaches, as recently done in the CAMEL approach [44].

Material and Methods

The principle of minimizing the change in enzyme usage between an alternative and reference state

To find the enzyme distribution vector that matches the enzyme usage of a cell growing in alternative growth conditions, we propose PARROT, an approach that minimizes the distance between a reference enzyme allocation $\mathbf{E_{ref}}$ and an alternative growth enzyme allocation $\mathbf{E_s}$ (Figure 1). This is consistent with observations that micro-organisms minimize expenditures to perform growth and to maintain the associated flux state [26]. We define and compare four different objectives to model the distance between the allocation of enzymes in alternative and reference growth states: (i) the Manhattan distance; (ii) the Euclidean distance; (iii) the weighted sum of the Manhattan distance between enzyme allocations and the Manhattan distance between flux distributions; (iv) the weighted sum of the Euclidean distance between enzyme allocations and the

Euclidean distance between flux distributions. The first can be formulated as a linear optimization problem (LP1), specified as follows:

$$\min \left\| \frac{\mathbf{E_{ref}}}{E_{ref}^{tot}} - \frac{\mathbf{E_s}}{E_s^{tot}} \right\|_1 \quad (2)$$

$$\text{s.t. } \mathbf{Nv} = \mathbf{0} \quad (3)$$

$$v_{s,min} \leq v_s \leq v_{s,max} \quad (4)$$

$$v_s \leq k_{cat} \cdot [E_s] \quad (5)$$

$$\sum E_s = E_s^{tot} \quad (6)$$

$$v_{bio} = \mu, \quad (7)$$

where E_{ref}^{tot} and E_s^{tot} represent the total enzyme usage in the model for the reference and alternative states, respectively; $\mathbf{N}$ is the stoichiometric matrix; $\mathbf{v}$ is the flux distribution vector; v_{bio} is the flux through the biomass pseudo-reaction; and μ is the specific growth rate, determined from measurements in the alternative state. The other objectives are captured by the following:

$$\text{QP1: } \left\| \frac{\mathbf{E_{ref}}}{E_{ref}^{tot}} - \frac{\mathbf{E_s}}{E_s^{tot}} \right\|_2, \quad (8)$$

$$\text{LP2: } \left\| \frac{\mathbf{E_{ref}}}{E_{ref}^{tot}} - \frac{\mathbf{E_s}}{E_s^{tot}} \right\|_1 + \lambda \|\mathbf{v_{ref}} - \mathbf{v_s}\|_1, \quad (9)$$

$$\text{QP2: } \left\| \frac{\mathbf{E_{ref}}}{E_{ref}^{tot}} - \frac{\mathbf{E_s}}{E_s^{tot}} \right\|_2 + \lambda \|\mathbf{v_{ref}} - \mathbf{v_s}\|_2. \quad (10)$$

where the parameter λ is a weighting factor chosen by inspecting the difference between the norms of enzyme allocation and the flux distributions. We solved the corresponding problems under the same constraints as in Eq. 2. We implemented and solved the problems in MATLAB (The MathWorks Inc., Natick, Massachusetts) using the COBRA Toolbox [45] and the Gurobi solver v9.1.1 [46]. The implementation of PARROT can be found in the GitHub repository: https://github.com/mauricioamf/PARROT.

Experimental data and simulation constraints

To test the variants of the proposed approach, PARROT, we used the pcGEMs of *Saccharomyces cerevisiae*, ecYeast8 [47], and *Escherichia coli*, eciML1515 [48]. We employed quantitative proteomics measurements for both species performed in a number of growth conditions, ranging from optimal growth in standard physiological conditions to stress conditions, alternative nutrient usage and chemostat cultivation.

For *S. cerevisiae*, we used the protein measurements from Chen and Nielsen [23] for 19 different growth conditions, which were collected from four studies [19–22]. These included proteomics measurements in yeast growing in ethanol, osmolarity, and high temperature stresses [19]; yeast growing in chemostats with reducing nitrogen availability [20]; and yeast growing in chemostats limited by the nitrogen source in increasing dilution rates and in chemostats with alternative nitrogen sources [22]. We also used measurements of nutrient uptake rates, growth rates and protein content from these studies to constrain the batch model, which does not consider protein measurements and rely on the protein pool constraint.

For *E. coli*, we used the proteomics data for 20 different growth conditions collected in [50] from three different studies [23–25]. These include batch cultivations of *E. coli* growing with different carbon sources and a glucose-limited chemostat culture, with dilution rates ranging from 0.12 h^{-1} to 0.5 h^{-1} performed by Schmidt et al. [25], a second chemostat limited by glucose at dilution rates ranging from 0.11 h^{-1} to 0.49 h^{-1} [23], and a third chemostat limited by glucose at dilutions rates ranging from 0.21 h^{-1} to 0.51 h^{-1} [24]. Similar to *S. cerevisiae*, the batch model was constrained with the nutrient uptake rates, growth rates and protein content measured in the studies where the protein measurements were taken. For both species, we excluded the conditions that did not have measured uptake rates, growth rates, or protein content. In addition, we excluded the temperature stress conditions from Lahtvee et al. [19], as temperature can severely impact the function of enzymes [51], and temperature stress responses entail changes beyond metabolic flux redistribution [17]. To prevent overconstraining the models, we allowed 5% flexibility on the growth rate. For other constraints, we allowed flexibility by increments of 1% until the measured growth rate is achieved.

Pre-processing of protein measurements for the reference state

From the protein measurements obtained from Davidi et al. [50] and Chen and Nielsen [49] we separated the measurements according to each experiment performed in the original studies. From each experiment we selected the control sample to represent the reference state in our approach PARROT. We corrected the protein measurements for the reference state measurements by integrating the values into the pcGEMs ecYeast8 and eciML1515 for *S. cerevisiae* and *E. coli*, respectively, using the GECKO Toolbox 2 [48]. The GECKO Toolbox 2 identifies the enzyme usage values that most

limit growth and allow flexibility in the enzyme usage constraints to prevent over-constraining the model. For the $\mathbf{E_{ref}}$ vector of each experiment, we then used these relaxed protein measurements, while keeping the original values for proteins that were unchanged by allowing flexibility.

Assessment of predicted enzyme usage distributions

The protein measurements, $\mathbf{E_s^{exp}}$, for the alternative growth conditions obtained from Davidi et al. [50] and Chen and Nielsen [49] were not used directly in simulations. These experimental measurements were instead employed to calculate a baseline to which predictions of $\mathbf{E_s}$ were compared. Assuming that simulations performed with pcGEMs use only the optimal concentration of enzymes necessary to carry a given metabolic flux, the model-allocated protein usage would underestimate the *in vivo* enzyme concentrations. To allow for a fair comparison, we devised a baseline by integrating the experimental proteomics measurements of each experiment into the pcGEMs using the GECKO Toolbox 2 in which we minimized the total enzyme allocation given the following optimization problem:

$$\min \left\| \mathbf{E_s^{exp}} \right\|_1 \quad (11)$$

$$\text{s.t. } \mathbf{Nv} = \mathbf{0} \quad (12)$$

$$\mathbf{v_{s,min}} \leq \mathbf{v_s} \leq \mathbf{v_{s,max}} \quad (13)$$

$$v_{s,j} \leq k_{cat}^{ij} \cdot \left[E_s^{exp,i} \right] \quad (14)$$

$$\sum E_s^{exp} = E_s^{exp,tot} \quad (15)$$

$$v_{bio} = \mu. \quad (16)$$

The resulting enzyme usage distribution, $\mathbf{E_s^{exp}}$, was then defined as the baseline for each sample of each proteomics experiment. We compared the predicted $\mathbf{E_s}$ values from the four variants of PARROT to $\mathbf{E_s^{exp}}$ by calculating the Pearson correlations of each sample. Further, we calculated the root-median square error (RMdSE) to measure the difference between predicted and baseline values. For assessing both correlations and the RMdSE, we log10-transformed the values for the predictions and the baseline.

We also performed a robustness analysis by checking the effect of minimizing the 2-norm of $\mathbf{E_s^{exp}}$ to construct the baseline, instead of the 1-norm, keeping the constraints defined in Equations 12-16:

$$\min \left\| \mathbf{E_s^{exp}} \right\|_2 \quad (17)$$

We compared the predictions of our approaches to those obtained using an extension of parsimonious enzyme usage FBA (pFBA) [52] to consider enzyme constraints. It is suitable for benchmarking our approach given its ability to predict phenotypes in good accordance with other methods that rely on transcriptomics and proteomics data. To this end, for each sample of each experiment, we defined the optimization problem as:

$$\min \sum_{j=1}^{m} v_{j,s,irrev} \quad (18)$$

$$\text{s.t. } \mathbf{N_{s,irrev}} \cdot \mathbf{v_{s,irrev}} = \mathbf{0} \quad (19)$$

$$0 \leq \mathbf{v_{s,irrev}} \leq \mathbf{v_{s,irrev,max}} \quad (20)$$

$$v_{s,irrev,j} \leq k_{cat}^{ij} \cdot [E_{s,i}] \quad (21)$$

$$\sum E_s = E_s^{tot} \quad (22)$$

$$v_{bio} = \mu, \quad (23)$$

where $v_{j,s,irrev}$ corresponds to the flux distribution of an irreversible model in an alternative growth condition. We also assessed a modified version of pFBA with enzyme constraints with the following objective:

$$\min \sum_{j=1}^{m} E_{s,i} \cdot k_{cat}^{ij} . \quad (24)$$

For pFBA and the modified implementation, we applied the same constraints on nutrient uptake rates and growth rates as for the four approaches assessed previously, and calculated the Pearson correlations and the RMdSE. Lastly, as a negative control to benchmark the performance of PARROT, we equated $E_{s,i}$ to k_{cat}^{ij} ($E_{s,i} = k_{cat}^{ij}$), meaning that k_{cat} values we used directly as the enzyme usage. This allows for determining how much of the correlation between $\mathbf{E_{exp}}$ and $\mathbf{E_s}$ can be attributed directly to k_{cat} values. We calculated the correlation values and RMdSE for all assessed optimization problems and compared them to the predictions of pFBA and its modified implementation using a Pairwise Wilcoxon rank sum test with Bonferroni correction.

Assessment of optimal values for the λ weighting factor

To systematically assess the impact of different lambda values, we optimised the LP2 and QP2 variants using λ values ranging from 0 (no fluxes used) to 1 (fluxes and enzyme usages equally considered). Additionally, we optimised the LP2 and QP2

variants using λ values ranging from 0.1 to 1 in order to make sure fluxes are always used for the objective function. In both scenarios, we calculated the Pearson correlation to the baseline for each λ value. We determined the optimal λ value as the value that outputs predictions with the highest Pearson correlation when compared to the first norm baseline.

Acknowledgements

We thank Marius Arend, Philipp Wendering and Eduardo Almeida for their critical discussion and comments on this study.

Data availability

All data and code are publicly available in the GitHub repository: (https://github.com/mauricioamf/PARROT)

References

1. Price ND, Papin JA, Schilling CH, Palsson BO. Genome-scale microbial in silico models: The constraints-based approach. Trends Biotechnol. 2003;21: 162–169. doi:10.1016/S0167-7799(03)00030-1
2. 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. Price ND, editor. PLoS Comput Biol. 2012;8: e1002575. doi:10.1371/journal.pcbi.1002575
3. 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: 935. doi:10.15252/msb.20167411
4. Bulik S, Holzhütter HG, Berndt N. The relative importance of kinetic mechanisms and variable enzyme abundances for the regulation of hepatic glucose metabolism - insights from mathematical modeling. BMC Biol. 2016;14: 1–22. doi:10.1186/S12915-016-0237-6/FIGURES/18

5. Basan M, Hui S, Okano H, Zhang Z, Shen Y, Williamson JR, et al. Overflow metabolism in Escherichia coli results from efficient proteome allocation. Nature. 2015;528: 99–104. doi:10.1038/nature15765
6. 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: 12663–12668. doi:10.1073/pnas.0609845104
7. Bekiaris PS, Klamt S. Automatic construction of metabolic models with enzyme constraints. BMC Bioinformatics. 2020;21: 1–13. doi:10.1186/S12859-019-3329-9/TABLES/2
8. Ferreira MA de M, Silveira WB da, Nikoloski Z. Protein constraints in genome-scale metabolic models: data integration, parameter estimation, and prediction of metabolic phenotypes. Authorea Preprints. 2022. doi:10.22541/AU.166082043.36599845/V1
9. Lindemann C, Thomanek N, Hundt F, Lerari T, Meyer HE, Wolters D, et al. Strategies in relative and absolute quantitative mass spectrometry based proteomics. Biol Chem. 2017;398: 687–699. doi:10.1515/HSZ-2017-0104/ASSET/GRAPHIC/J_HSZ-2017-0104_FIG_005.JPG
10. Otto A, Becher D, Schmidt F. Quantitative proteomics in the field of microbiology. Proteomics. 2014;14: 547–565. doi:10.1002/pmic.201300403
11. Calderón-Celis F, Encinar JR, Sanz-Medel A. Standardization approaches in absolute quantitative proteomics with mass spectrometry. Mass Spectrom Rev. 2018;37: 715–737. doi:10.1002/mas.21542
12. Swiatly A, Plewa S, Matysiak J, Kokot ZJ. Mass spectrometry-based proteomics techniques and their application in ovarian cancer research. J Ovarian Res. 2018;11: 1–13. doi:10.1186/s13048-018-0460-6
13. Li H, Siddiqui O, Zhang H, Guan Y. Joint learning improves protein abundance prediction in cancers. BMC Biol. 2019;17: 1–14. doi:10.1186/S12915-019-0730-9/FIGURES/6
14. Terai G, Asai K. Improving the prediction accuracy of protein abundance in Escherichia coli using mRNA accessibility. Nucleic Acids Res. 2020;48. doi:10.1093/nar/gkaa481

15. Ferreira M, Ventorim R, Almeida E, Silveira S, Silveira W. Protein Abundance Prediction Through Machine Learning Methods. J Mol Biol. 2021;433: 167267. doi:10.1016/J.JMB.2021.167267
16. Heckmann D, Lloyd CJ, Mih N, Ha Y, Zielinski DC, Haiman ZB, et al. Machine learning applied to enzyme turnover numbers reveals protein structural correlates and improves metabolic models. Nat Commun. 2018;9: 1–10. doi:10.1038/s41467-018-07652-6
17. Lahtvee P-J, Kumar R, Hallstrom BM, Nielsen J. Adaptation to different types of stress converge on mitochondrial metabolism. Mol Biol Cell. 2016;27: 2505–2514. doi:10.1091/mbc.E16-03-0187
18. Segrè D, Vitkup D, Church GM. Analysis of optimality in natural and perturbed metabolic networks. Proc Natl Acad Sci U S A. 2002;99: 15112–15117. doi:10.1073/PNAS.232349399/SUPPL_FILE/3493SUPPLINKS.HTML
19. Lahtvee PJ, Sánchez BJ, Smialowska A, Kasvandik S, Elsemman IE, Gatto F, et al. Absolute Quantification of Protein and mRNA Abundances Demonstrate Variability in Gene-Specific Translation Efficiency in Yeast. Cell Syst. 2017;4: 495-504.e5. doi:10.1016/j.cels.2017.03.003
20. Yu R, Campbell K, Pereira R, Björkeroth J, Qi Q, Vorontsov E, et al. Nitrogen limitation reveals large reserves in metabolic and translational capacities of yeast. Nat Commun. 2020;11: 1–12. doi:10.1038/s41467-020-15749-0
21. Di Bartolomeo F, Malina C, Campbell K, Mormino M, Fuchs J, Vorontsov E, 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;117: 7524–7535. doi:10.1073/PNAS.1918216117/SUPPL_FILE/PNAS.1918216117.SD07.XLSX
22. Yu R, Vorontsov E, Sihlbom C, Nielsen J. Quantifying absolute gene expression profiles reveals distinct regulation of central carbon metabolism genes in yeast. Elife. 2021;10. doi:10.7554/ELIFE.65722
23. Valgepea K, Adamberg K, Seiman A, Vilu R. Escherichia coli achieves faster growth by increasing catalytic and translation rates of proteins. Mol Biosyst. 2013;9: 2344–2358. doi:10.1039/C3MB70119K

24. Peebo K, Valgepea K, Maser A, Nahku R, Adamberg K, Vilu R. Proteome reallocation in Escherichia coli with increasing specific growth rate. Mol Biosyst. 2015;11: 1184–1193. doi:10.1039/c4mb00721b
25. Schmidt A, Kochanowski K, Vedelaar S, Ahrné E, Volkmer B, Callipo L, et al. The quantitative and condition-dependent Escherichia coli proteome. Nat Biotechnol. 2016;34: 104–110. doi:10.1038/nbt.3418
26. Goelzer A, Muntel J, Chubukov V, Jules M, Prestel E, Nölker R, et al. Quantitative prediction of genome-wide resource allocation in bacteria. Metab Eng. 2015;32: 232–243. doi:10.1016/J.YMBEN.2015.10.003
27. Navarro-Tapia E, Pérez-Torrado R, Querol A. Ethanol effects involve non-canonical unfolded protein response activation in yeast cells. Front Microbiol. 2017;8: 1–12. doi:10.3389/fmicb.2017.00383
28. Alvim MCT, Vital CE, Barros E, Vieira NM, da Silveira FA, Balbino TR, et al. Ethanol stress responses of Kluyveromyces marxianus CCT 7735 revealed by proteomic and metabolomic analyses. Antonie Van Leeuwenhoek. 2019;112: 827–845. doi:10.1007/s10482-018-01214-y
29. Silveira FA, de Oliveira Soares DL, Bang KW, Balbino TR, de Moura Ferreira MA, Diniz RHS, et al. Assessment of ethanol tolerance of Kluyveromyces marxianus CCT 7735 selected by adaptive laboratory evolution. Appl Microbiol Biotechnol. 2020;104: 7483–7494. doi:10.1007/s00253-020-10768-9
30. Javkar K, Rand H, Hoffmann M, Luo Y, Sarria S, Thirunavukkarasu N, et al. Whole-Genome Assessment of Clinical Acinetobacter baumannii Isolates Uncovers Potentially Novel Factors Influencing Carbapenem Resistance. Front Microbiol. 2021;12: 714284. doi:10.3389/FMICB.2021.714284/BIBTEX
31. Navon-Venezia S, Kondratyeva K, Carattoli A. Klebsiella pneumoniae: a major worldwide source and shuttle for antibiotic resistance. FEMS Microbiol Rev. 2017;41: 252–275. doi:10.1093/FEMSRE/FUX013
32. Moriya H. Quantitative nature of overexpression experiments. Mol Biol Cell. 2015;26: 3932–3939. doi:10.1091/MBC.E15-07-0512/ASSET/IMAGES/LARGE/MBC-26-3932-G004.JPEG
33. Kok M, Maton L, van der Peet M, Hankemeier T, van Hasselt JGC. Unraveling antimicrobial resistance using metabolomics. Drug Discov Today. 2022;27: 1774–1783. doi:10.1016/J.DRUDIS.2022.03.015

34. von der Haar T. A quantitative estimation of the global translational activity in logarithmically growing yeast cells. BMC Syst Biol. 2008;2: 1–14. doi:10.1186/1752-0509-2-87/FIGURES/7
35. O'Brien EJ, Utrilla J, Palsson BO. Quantification and Classification of E. coli Proteome Utilization and Unused Protein Costs across Environments. PLoS Comput Biol. 2016;12: e1004998. doi:10.1371/JOURNAL.PCBI.1004998
36. Yang Y, Karin O, Mayo A, Song X, Chen P, Santos AL, et al. Damage dynamics and the role of chance in the timing of E. coli cell death. Nature Communications 2023 14:1. 2023;14: 1–11. doi:10.1038/s41467-023-37930-x
37. Chen M, Xie T, Li H, Zhuang Y, Xia J, Nielsen J. Yeast increases glycolytic flux to support higher growth rates accompanied by decreased metabolite regulation and lower protein phosphorylation. Proceedings of the National Academy of Sciences. 2023;120: e2302779120. doi:10.1073/PNAS.2302779120/SUPPL_FILE/PNAS.2302779120.SD01.XLSX
38. Deparis Q, Claes A, Foulquié-Moreno MR, Thevelein JM. Engineering tolerance to industrially relevant stress factors in yeast cell factories. FEMS Yeast Res. 2017;17: 1–35. doi:10.1093/femsyr/fox036
39. Mori M, Schink S, Erickson DW, Gerland U, Hwa T. Quantifying the benefit of a proteome reserve in fluctuating environments. Nat Commun. 2017;8: 1–8. doi:10.1038/s41467-017-01242-8
40. Hackett SR, Zanotelli VRT, Xu W, Goya J, Park JO, Perlman DH, et al. Systems-level analysis of mechanisms regulating yeast metabolic flux. Science. 2016;354. doi:10.1126/SCIENCE.AAF2786
41. Lalanne JB, Taggart JC, Guo MS, Herzel L, Schieler A, Li GW. Evolutionary Convergence of Pathway-Specific Enzyme Expression Stoichiometry. Cell. 2018;173: 749-761.e38. doi:10.1016/J.CELL.2018.03.007
42. Alter TB, Blank LM, Ebert BE. Proteome Regulation Patterns Determine Escherichia coli Wild-Type and Mutant Phenotypes. mSystems. 2021;6. doi:10.1128/msystems.00625-20
43. Kerkhoven EJ. Advances in constraint-based models: methods for improved predictive power based on resource allocation constraints. Curr Opin Microbiol. 2022;68: 102168. doi:10.1016/J.MIB.2022.102168

44. Ferreira MA de M, Wendering P, Arend M, Silveira WB da, Nikoloski Z. Accurate prediction of in vivo protein abundances by coupling constraint-based modelling and machine learning. bioRxiv. 2023; 2023.04.11.536445. doi:10.1101/2023.04.11.536445
45. Heirendt L, Arreckx S, Pfau T, Mendoza SN, Richelle A, Heinken A, et al. Creation and analysis of biochemical constraint-based models using the COBRA Toolbox v.3.0. Nat Protoc. 2019;14: 639–702. doi:10.1038/s41596-018-0098-2
46. Gurobi Optimization L. Gurobi Optimizer Reference Manual. 2020.
47. 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;10. doi:10.1038/s41467-019-11581-3
48. Domenzain I, Sánchez B, Anton M, Kerkhoven EJ, Millán-Oropeza A, Henry C, et al. Reconstruction of a catalogue of genome-scale metabolic models with enzymatic constraints using GECKO 2.0. Nat Commun. 2022;13: 1–13. doi:10.1038/s41467-022-31421-1
49. Chen Y, Nielsen J. In vitro turnover numbers do not reflect in vivo activities of yeast enzymes. Proc Natl Acad Sci U S A. 2021;118: e2108391118. doi:10.1073/PNAS.2108391118/SUPPL_FILE/PNAS.2108391118.SD08.XLSX
50. Davidi D, Noor E, Liebermeister W, Bar-Even A, Flamholz A, Tummler K, et al. Global characterization of in vivo enzyme catalytic rates and their correspondence to in vitro kcat measurements. Proc Natl Acad Sci U S A. 2016;113: 3401–3406. doi:10.1073/pnas.1514240113
51. Wendering P, Nikoloski Z. Model-driven insights into the effects of temperature on metabolism. Biotechnol Adv. 2023;67: 108203. doi:10.1016/J.BIOTECHADV.2023.108203
52. 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. doi:10.1038/msb.2010.47

3.3 Accurate prediction of *in vivo* protein abundances by coupling constraint-based modelling and machine learning

Text adapted from the unmarked revised manuscript as accepted by the Metabolic Engineering journal, available at: https://doi.org/10.1016/j.ymben.2023.09.014.

Abstract

Quantification of how different environmental cues affect protein allocation can provide important insights for understanding cell physiology. While absolute quantification of proteins can be obtained by resource-intensive mass-spectrometry-based technologies, prediction of protein abundances offers another way to obtain insights into protein allocation. Here we present CAMEL, a framework that couples constraint-based modelling with machine learning to predict protein abundance for any environmental condition. This is achieved by building machine learning models that leverage static features, derived from protein sequences, and condition-dependent features predicted from protein-constrained metabolic models. Our findings demonstrate that CAMEL results in excellent prediction of protein allocation in *E. coli* (average Pearson correlation of at least 0.9), and moderate performance in *S. cerevisiae* (average Pearson correlation of at least 0.5). Therefore, CAMEL outperformed contending approaches without using molecular read-outs from unseen conditions and provides a valuable tool for using protein allocation in biotechnological applications.

Keywords: Protein allocation, Environmental effects, Multi-model framework

Introduction

Proteomics technologies facilitate system-wide profiling of the identity as well as changes in abundance, distribution, modifications, and interactions of proteins that drive different cellular processes (Schubert et al., 2017). As cells are dynamic systems, these characteristics of the proteome change in response to different environmental cues to facilitate the system's adaptation to different physiological states (Liu et al., 2019; Nielsen, 2019). As a result, high-throughput quantitative proteomics technologies, based on mass spectrometry (MS) techniques, have provided valuable

information for various biotechnological and medical applications (Kültz, 2020; Lill et al., 2021). However, the measurement of absolute protein abundance still poses great challenges, due to physicochemical properties that interfere with ionization efficiency in mass spectrometry (Otto et al., 2014; Pappireddi et al., 2019) and the lack of a standardized approach, affecting the reproducibility of resulting data (Calderón-Celis et al., 2018).

Another approach for protein quantification relies on predicting protein abundances by using a variety of machine learning approaches using features that are more facile to measure or quantify. For instance, with gene expression data as predictors, Torres-García et al. (2009) trained gradient-boosted trees to predict protein abundances in *Desulfovibrio vulgaris*, resulting in prediction accuracies, quantified by the coefficient of determination, R^2, ranging from 0.39 to 0.58. Similarly, Mehdi et al. (2014) used Bayesian networks to predict protein abundance for *Saccharomyces cerevisiae* and *Schizosaccharomyces pombe* using gene expression data, yielding Pearson's correlation coefficients ranging from 0.50 to 0.71. Mergner et al. (2020) used stepwise and LASSO regression models to predict protein abundances in *Arabidopsis thaliana* based on gene expression as well as sequence-derived features, achieving Pearson correlation in the range from 0.61 to 0.79 across different tissues. Codon usage metrics have also been used to predict protein abundance in *Saccharomyces cerevisiae*, with the best model achieving an R^2 value of 0.74 (Ferreira et al., 2021). In addition, structural features of mRNA molecules have been used to predict protein abundances in *Escherichia coli*, achieving a Spearman correlation coefficient of 0.71. However, these machine learning models have been developed using quantitative proteomics and transcriptomics data from optimal growth conditions and their performance in sub-optimal growth conditions remains unexplored. In addition, given the dynamic nature of the proteome, it is expected that the predictions across different conditions from machine learning models based on environment-invariant features (e.g., codon usage and structure-related), are poor.

The advent of protein-constrained genome-scale metabolic models (pcGEMs) (Beg et al., 2007) has facilitated not only the prediction of protein abundance but also the usage of proteomics data to predict metabolic and physiological traits (Bekiaris and Klamt, 2020; Domenzain et al., 2022; Sánchez et al., 2017). This is achieved by parameterizing pcGEMs with data on enzyme turnover numbers, providing the basis

to link steady-state fluxes with enzyme abundances. For instance, a pcGEM of *E. coli* was used by Adadi et al. (2012) to predict protein abundance, achieving Pearson correlation coefficient of 0.84 with gene expression data of *E. coli* grown on a glucose minimal medium. However, given that comparisons were made with gene expression data, it remains unexplored how these predictions fare against measured proteomics data. Further, Heckmann et al. (2018) trained machine learning approaches to predict enzyme catalytic efficiency, namely *in vitro* (k_{cat}) and *in vivo* values (k_{max}^{vivo}), that were later used to predict protein abundances with the approach of Adadi et al. (2012). Comparing the predicted protein abundances to experimental data from Schmidt et al. (2016), they found that using k_{max}^{vivo} values resulted in a 43% lower root mean squared error (RMSE) in comparison to k_{cat} values. However, when comparing the prediction to measured proteomics data, the obtained R^2 values were poor to modest, ranging from 0.11 to 0.68. As a result, we still lack an understanding of the factors that affect the quality of the protein abundance predictions. Moreover, like the machine learning approaches above, the existing predictions of protein abundance based on pcGEMs have only been assessed under conditions that achieve maximal measured specific growth rate for the wild-type strain that we refer to as optimal. Therefore, the problem of predicting protein abundance under sub-optimal, stress-related conditions, that result in smaller specific growth rate than the optimal, is particularly difficult since many mechanisms affecting protein abundance (e.g., codon usage (Novoa et al., 2019), RNA levels (Eraslan et al., 2019), protein interactions (Mergner et al., 2020), 3' or 5' UTR motifs (Terai and Asai, 2020)) are not included in pcGEMs.

To address these issues, we coupled constraint-based modelling with machine learning to predict *in vivo* protein abundances, leading to a multi-model framework termed CAMEL (Coupled Approach of MEtabolic modelling and machine Learning) that leverages the strengths of both approaches. First, CAMEL predicts the protein abundance in different conditions using a constraint-based approach. Next, these protein abundance predictions are employed together with experimentally measured protein abundances to calculate protein reserve ratios. CAMEL then trains machine learning (ML) models to predict the protein reserve ratios using different static features, derived from the protein sequence, and condition-dependent features, obtained from the pcGEM. Lastly, the predicted protein reserve ratios along with the predicted protein abundances from the pcGEM are used to calculate *in vivo* protein abundances. Once

trained, the models of CAMEL can be used to make predictions by using features predicted by the pcGEM instead of data on molecular read-outs from unseen conditions. Our results showed that the CAMEL approach outperformed contending methods on data from two model organisms, *E. coli* and *S. cerevisiae*, highlighting the advantages of a coupling of constraint-based modelling and machine learning approaches. In addition, these applications of CAMEL point to particular proteins and metabolic pathways for which protein abundance is difficult to predict, requiring the consideration of additional features to improve prediction performance in future developments.

Material and Methods

Data set construction

We obtained the experimental enzyme abundances for *E. coli* over 31 growth conditions from three different studies (Peebo et al., 2015; Schmidt et al., 2016; Valgepea et al., 2013), as reported in Davidi et al. (2016). The growth conditions in these experiments range from alternative carbon substrates (e.g., acetate, glucosamine, glucose, glycerol, mannose, pyruvate, and xylose) to chemostats with increasing dilution rates, with glucose as a carbon source. For *S. cerevisiae*, we used the enzyme abundance values from the data set of Chen and Nielsen (2021), which contains quantitative proteomics measurements for 30 growth conditions obtained from four studies (Chen and Nielsen, 2021; Di Bartolomeo et al., 2020; Lahtvee et al., 2017; Yu et al., 2020), including: optimal batch growth with glucose as carbon source, nitrogen- and glucose-limited chemostats with increasing dilution rates, stress states achieved with high ethanol and osmolarity, and reduced nutrient availability.

Condition-dependent features are predicted from the pcGEMs and include the enzyme usage and the metabolic fluxes associated with each enzyme i. These features are obtained by solving a linear problem to minimize excess enzyme usage (Eqs. (1)-(8)), detailed below. To predict the enzyme usage distributions, we used the pcGEMs eciML1515 (Domenzain et al., 2022) and ecYeast8 (Lu et al., 2019), from *E. coli* and *S. cerevisiae*, respectively, both without integrated enzyme measurements (batch model). To this end, we minimized excess enzyme usage. Maximal enzyme usage is linked to efficient usage of resources, which has been shown to result in better estimates of k_{cat} values over other approaches that focus on minimization of total flux

(Xu et al., 2021). This objective function also ensures that cellular burden is alleviated, which can lead to proteotoxic effects for the cell (Kintaka et al., 2020). We solve the following linear programming (LP) problem, termed LP1:

$$\min_{\mathrm{v,E}} E_{s,tot} - \sum_{i=1}^{p} MW_i \cdot E_{s,i} \quad (1)$$

s. t.

$$\mathbf{N} \cdot \mathbf{v} = \mathbf{0} \quad (2)$$

$$v_{min,j} \leq v_j \leq v_{max,j}, 1 \leq j \leq r \quad (3)$$

$$v_{uptake,s} = v_{uptake,s}^{exp} \quad (4)$$

$$v_{secretion,s} = v_{secretion,s}^{exp} \quad (5)$$

$$v_{bio} = \mu_s \quad (6)$$

$$v_j \leq k_{cat}^{ij} \cdot E_{s,i}, 1 \leq j \leq r, 1 \leq i \leq p \quad (7)$$

$$E_{s,tot} = P_{s,tot}^{exp} \cdot f \cdot \sigma \quad (8)$$

where $E_{s,i}$ is the enzyme i, ranging from 1 to the number of proteins p in the model ($1 \leq i \leq p$), in condition s, MW_i is the molecular weight of an enzyme i, $\mathbf{N}$ is the stoichiometric matrix, $\mathbf{v}$ is the flux distribution vector, v_j is the metabolic flux through reaction j, ranging from 1 to the number of reactions r in the model ($1 \leq j \leq r$), v_{bio} is flux through the biomass reaction, and μ is the specific growth rate, E_{tot} is the total enzyme content, P_{tot}^{exp} is the total protein content, f is the mass fraction of all measured proteins included in the model, and σ is the average *in vivo* enzyme saturation, assumed to be a specific value of 0.5 for conditions in which this parameter is not available (Domenzain et al., 2022; Sánchez et al., 2017). The constraints applied to the problem are based on the available data about specific growth rates, secretion, and nutrient uptake rates, protein usage and total protein content.

For both organisms, we constrained the models given the measurements of nutrient uptake rates and specific growth rates using data from Davidi et al. (2016) and Chen and Nielsen (2021). We excluded the experimental conditions that lacked physiological data (e.g., nutrient uptake rates, specific growth rates, or protein content). In addition, we opted to exclude the temperature stress conditions from Lahtvee et al. (2017), as temperature stress triggers responses that entail cellular changes that can

impact the function of enzymes (Li et al., 2021) and require more tailored modelling approaches (Wendering et al., 2023). This filtering resulted in data from 20 conditions for *E. coli*, and 19 conditions for *S. cerevisiae*. For conditions for which the models were overconstrained, we relaxed the constraints on the uptake and secretion rates by increments of 1% until the measured specific growth rate could be achieved.

We were also interested in whether using pcGEMs integrated with *in vivo* catalytic rates (k_{max}^{vivo}) instead of *in vitro* values (k_{cat}) would impact the performance of the ML models. To this end, we exchanged the integrated k_{cat} values with the k_{max}^{vivo} values, obtained from the compilation by (Wendering et al., 2023) of estimated catalytic rates using pFBA from the Davidi et al. (2016) and Chen and Nielsen (2021) studies. For both *E. coli* and *S. cerevisiae*, we concatenated the predictions generated from the considered growth conditions into a single data set for each organism. The resulting data set for *E. coli* included 2256 enzymes for the pcGEM using k_{cat} values, and a data set with 2246 enzymes for the pcGEM k_{max}^{vivo} values. For *S. cerevisiae*, the resulting data set included 4596 enzymes for the pcGEM using k_{cat} values, and a data set with 4590 enzymes for the pcGEM k_{max}^{vivo} values.

Training and assessment of machine learning models

To train the machine learning models, we used the Tree-based Pipeline Optimization Tool (TPOT), an automated machine learning tool that optimizes machine learning pipelines to predict the target variable using genetic programming (Le et al., 2020). We configured TPOT for regression problems since we aimed to obtain quantitative predictions of the protein reserve ratio, ϕ_i, for each enzyme $i, 1 \leq i \leq p$. For all conditions in each species, we iterated TPOT over 100 generations to identify the optimized pipeline. We used a population size of 50, which is the number of the best pipelines that are predicted by TPOT in one generation that are then carried to next generation, but with randomly altered parameters in their constituent ML algorithms. We separated 80% of the constructed data sets for training and 20% for validation. The training data subset was used through all steps, while the validation data subsets were kept out of the model during TPOT optimization to prevent data leakage and/or training bias. For each training run, we further applied 10-fold cross-validation to the training data subset. All condition-dependent features were $\log_{10}$-transformed before training.

For performance assessment, we employed the adjusted coefficient of determination score (R^2_{adj}). It is often not possible to retrieve feature importance from TPOT-optimized pipelines, since many of the scikit-learn functions selected by TPOT do not support ranking of feature importance. As a result, we assessed how the condition-dependent features impact the predictions by removing them from the data sets and assessing how the TPOT-optimized pipelines of each data set perform. To further assess the prediction results, we used ϕ and E_s to obtain predictions about *in vivo* enzyme abundances $E_s^{exp\prime}$ which were compared with the experimental value using the Pearson's correlation coefficient. We performed GO enrichment analysis using the clusterProfiler R package (Wu et al., 2021; Yu et al., 2012). Further, we performed flux variability analysis (FVA) to assess the variability of central metabolic pathways and of enzyme usage pseudo-reactions.

Validation of models on genetically modified strains

To assess the validity of CAMEL on unseen conditions and its usefulness for metabolic engineering, we predicted the *in vivo* enzyme abundances $E_s^{exp\prime}$ of *E. coli* strains subjected to knock-out mutations followed by adaptive laboratory evolution (ALE) (McCloskey et al., 2018a, 2018b, 2018c, 2018d). For the optimization problem, we replicated the growth conditions by constraining the pcGEM eciML1515 using the estimated k_{max}^{vivo} values obtained by Heckmann et al., (2020), which were calculated from quantitative proteomics experiments on the same conditions as McCloskey et al. (2018a, 2018b, 2018c, 2018d). Next, we used the predicted E_s to predict ϕ using the machine learning models trained with either the k_{cat}-parameterized dataset or the k_{max}^{vivo}-parameterized dataset. We used the predicted ϕ and E_s to obtain calculate the *in vivo* enzyme abundances $E_s^{exp\prime}$ which we then compared with the experimental measurements of Heckmann et al., (2020) using the Pearson's correlation coefficient.

Results and Discussion

Machine learning accurately predicts protein reserve ratios

Our first observation is that the predicted abundance, $E_{s,i}$, of an enzyme i in condition s from pcGEMs is usually smaller than the measured *in vivo* abundance $E_{s,i}^{exp}$. This is the case since predictions on $E_{s,i}$ must match the corresponding flux, as $v_j = k_{cat}^{i,j} \cdot E_{s,i}$,

and $E_{s,i}$ cannot exceed the measured protein abundance. We term this discrepancy between the measured and predicted protein abundance for an enzyme i in condition s the protein reserve ratio, $\phi_{s,i}$, which is calculated by $\frac{E_{s,i}}{E_{s,i}^{exp}}$.

Since $\phi_{s,i}$ cannot be predicted from pcGEMs alone, we rely on machine learning to train models for ϕ_i, given data on measured protein abundance and predicted protein usage from pcGEMs over multiple conditions (Figure 1).

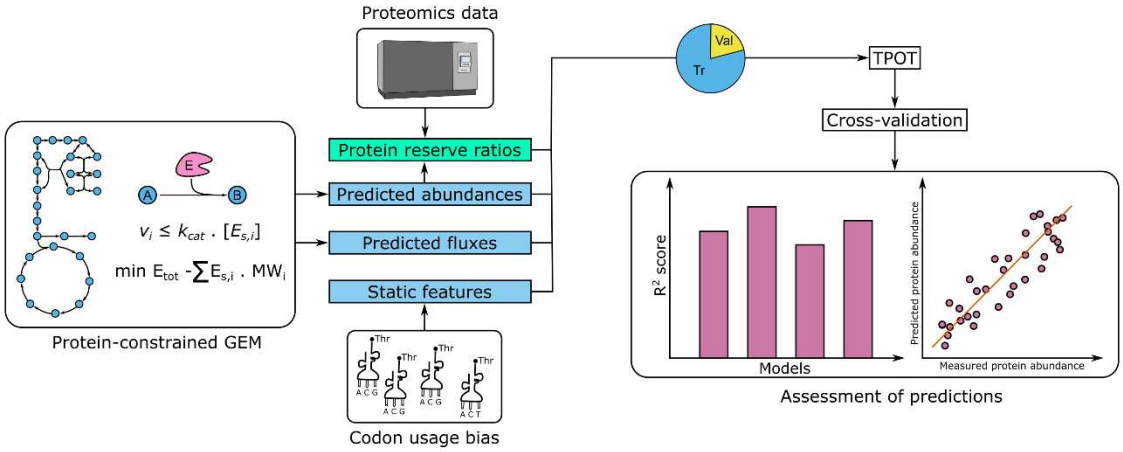

Figure 1. Schematic overview of CAMEL, a constraint-based approach to predict protein abundance. CAMEL uses two types of features, marked as blue boxes: static, obtained from codon usage metrics, and condition-dependent, given by metabolic flux and enzyme usage predicted by constraint-based modelling with pcGEMs. The predicted enzyme usage is employed together with experimental proteomics data to calculate protein reserve ratios that are then used as the target variable for machine learning (green box). The predictive model is obtained by selecting the optimal ensemble of machine learning approaches using TPOT. The performance is assessed by using cross-validation based on regression metrics. Abbreviations: Tr – training; Val – validation.

We refer to the collection $\boldsymbol{\phi}$ of reserve ratios over all proteins present in a pcGEM and measured by quantitative proteomic techniques as the distribution of **protein reserve ratios**. Since pcGEMs can be parameterized with *in vitro* measured or with *in vivo* estimated turnover numbers, denoted by k_{cat} and k_{max}^{vivo}, respectively, we performed a comparative analysis of protein reserve ratios in the two parameterization scenarios. Our first observation is that proteins differ with respect to their distributions of reserve ratios, which are dependent on the parameterization of the pcGEMs (Figure 2). For example, the protein reserve ratios of phosphoglycerate kinase in *E. coli* exhibit

a narrow distribution for the pcGEM parameterization based on k_{cat}, while the distribution is bimodal for the parameterization with k_{max}^{vivo}. In contrast, the phosphoglycerate kinase in *S. cerevisiae* has the same narrow distribution for both parameterizations. We also observed that the distributions of protein reserve ratios for all investigated enzymes have a heavy tail, due to very large reserve ratios in few conditions.

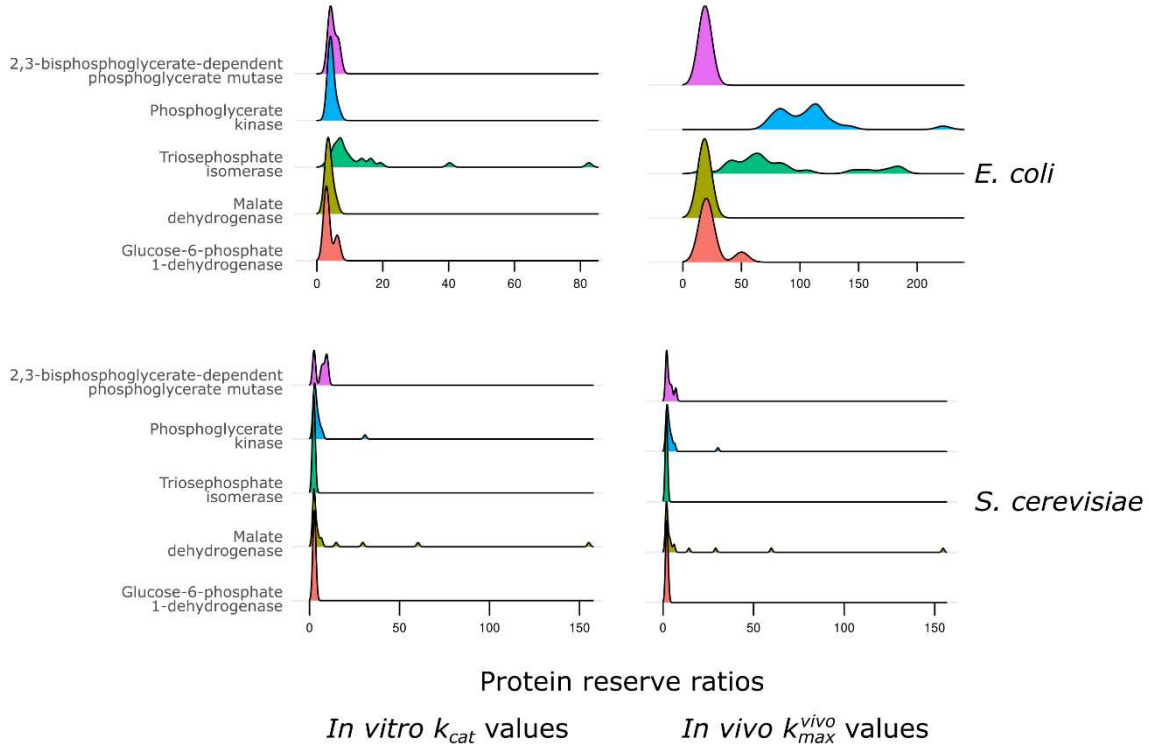

Figure 2. Distribution of protein reserve ratios. The selected proteins take part in central metabolic pathways such as glycolysis, pentose phosphate pathway and citric acid cycle, and are thus present in both organisms. Protein reserve ratios are unitless values.

Next, we used TPOT (Le et al., 2020) to optimize machine learning pipelines to predict **ϕ** based on two types of features—static and condition-dependent (Table S1). Static features were obtained from codon usage metrics, while condition-dependent features included enzyme usage and fluxes of the catalyzed reactions predicted by the pcGEM in the two parameterization scenarios (see Material and Methods). Our results indicated that the optimized pipelines for all considered cases exhibited similar structures, composed of two parts: starting with a stacked ensemble of linear and tree-based algorithms and then following with an Extreme Gradient Boosting (XGBoost)

algorithm (Chen and Guestrin, 2016), except for the *S. cerevisiae* model using k_{max}^{vivo} values, whose pipeline included the Extra-trees Regressor (Table S2). Using the trained and optimized pipelines, we then predicted $\boldsymbol{\phi}$ for the validation data set, which had not been used during model training and optimization. For *E. coli*, the machine learning models using either the k_{cat} or k_{max}^{vivo} values showed excellent performance (R_{adj}^2 > 0.9), while the model for *S. cerevisiae* showed worse performance (R_{adj}^2≈ 0.7) (Figure 3). By comparing the R_{adj}^2 values, we found that for *E. coli*, the pipelines with condition-dependent features derived by using k_{max}^{vivo} values outperformed those based on k_{cat} values by 6.7% (Figure 3). For yeast, however, we did not observe a difference in the overall performance of the machine learning models between the two parameterizations.

The reasons for the lower predictive performance for yeast in comparison to *E. coli* could be due to several factors: First, from a modelling perspective, there is more and better-curated knowledge about the metabolism of *E. coli* compared to that of *S. cerevisiae* (Bernstein et al., 2023). In addition, the metabolism of yeast is compartmentalized and includes mechanisms for enzyme regulation not accounted for in pcGEMs of any organism (e.g., organellar enzyme pools, post-translational modifications, diffusion effects). Second, from a data perspective, a lack of variability in the protein measurements, resulting from the consideration of samples from similar conditions, can hamper the training of machine learning models. Putting the two perspectives together, the eciML1515 model (1259 enzymes, (Domenzain et al., 2022)) includes a larger number of enzymes in comparison to the ecYeast8 model (965 enzymes, (Lu et al., 2019)), allowing to capture enzymes that may vary more across stress conditions. Lastly, the accuracy of the k_{cat} values in the BRENDA database (Chang et al., 2021) can also affect the quality of the predictions, given the challenges in measuring them experimentally.

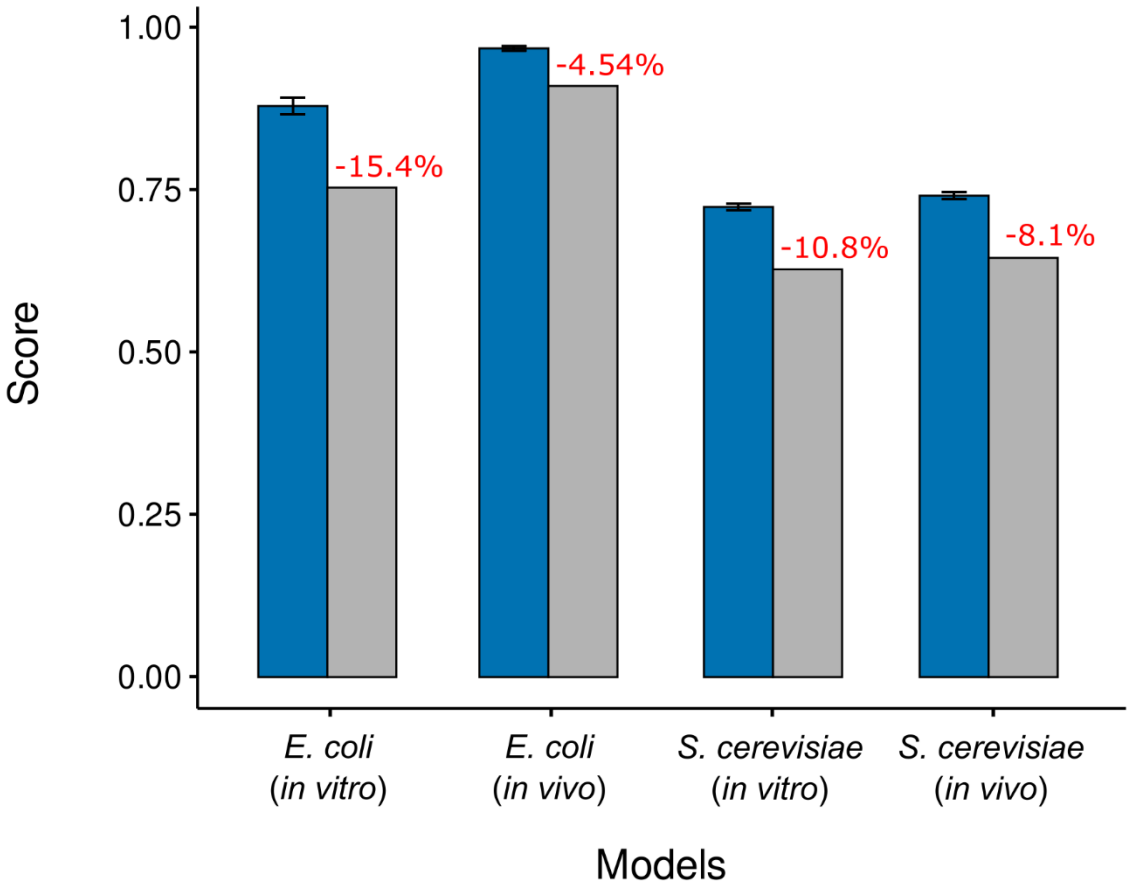

Figure 3. Performance of models from optimized pipelines. The figure shows a comparison of R^2_{adj} scores for the optimized pipelines based on all features (blue) or excluding condition-dependent features (grey) with either *in vitro* k_{cat} or k^{vivo}_{max} values. Negative percentages (red) indicate a decrease in predictive performance compared to scores obtained from the models trained using all features. Error bars indicate R^2_{adj} scores obtained from 10-fold cross-validation.

Importance of fluxes for predicting protein reserve ratios

We were also interested in identifying the extent to which fluxes for reactions associated with a protein affect the predictions of the corresponding protein reserve ratios. To this end, we removed these condition-dependent features from the training data set and re-trained the previously obtained TPOT-optimized pipelines using the reduced training data set. We found that the removal of fluxes resulted in reduced predictive performance in all considered scenarios. For instance, the obtained R^2_{adj} values for *E. coli* were at least 15.4% lower using k_{cat} values and 4.5% lower using k^{vivo}_{max}. For *S. cerevisiae*, there was a 10.8% decrease for the model using k_{cat} values

and an 8.1% decrease using k_{max}^{vivo} values (Figure 3). These findings demonstrated that the usage of fluxes as condition-dependent features improved model performance in both organisms, particularly in the case when pcGEMs were parametrized with k_{cat} values. In addition, this shows that the usage of fluxes as features may, to a certain extent, mitigate the effects of model parameterization.

The optimization problem of minimizing excess enzyme usage plays an important role in defining how the flux distribution and enzyme usage distribution are obtained. The objective function ensures efficient utilization of cellular resources, as it avoids burdening the cell with excess enzymes (Bruggeman et al., 2020). We emphasize that, biologically, excess enzymes are important for robustness and adaptability to changing environments, and its minimization could limit the flexibility needed by cells to adapt to these conditions, leading to reduced fitness (Alter et al., 2021). However, excess enzymes result in cellular burden, which can lead to proteotoxic effects for the cell, also reducing fitness (Kintaka et al., 2020). We emphasize that the protein reserve ratios we define do not necessarily correspond to biological protein reserves — we use this terminology to account for differences between predictions from pcGEMs and measured abundances. In that respect, one can use another objective to predict protein fraction used in fluxes; however, the key point is that we will still rely on ratio between measured and predicted values to build ML models — which is the essence of CAMEL.

Proteins with particular prediction error of protein reserve ratio are enriched in different pathways

Next, we were interested if there are any distinguishing characteristics between the two groups of proteins -- with low or high relative error in predictions. We considered a protein to show low prediction error if the relative error between the predicted and measured protein reserve ratio was lower than 50%, and a high prediction error if the relative error was higher than 100% in each model (noting that varying the definitions with lower, or respectively higher, values have a negligible effect on the conclusions).

First, we checked whether the two groups of proteins were enriched in any specific GO terms (Ashburner et al., 2000; Carbon et al., 2021). For *E. coli*, we found that proteins with low prediction error were enriched in biosynthesis of secondary metabolites, biosynthesis of cofactors, terpenoid biosynthesis and riboflavin

metabolism for models using either k_{cat} and k_{max}^{vivo} values (Figure S1). For proteins with high prediction error, enriched GO terms shared between models using either k_{cat} or k_{max}^{vivo} values included biosynthesis of secondary metabolites, carbon metabolism, fatty acid metabolism and degradation (Figure S1). The five proteins with the highest relative error (Table S3) for both model parameterizations were part of the same metabolic pathways, namely: glycolysis, fatty acid biosynthesis, beta-oxidation, and cell wall biogenesis.

Similarly, for *S. cerevisiae*, the proteins with low prediction error for the model using k_{cat} and k_{max}^{vivo} values were enriched in the biosynthesis of secondary metabolites, biosynthesis of amino acids, and carbon metabolism. Likewise, proteins with high prediction error were enriched in biosynthesis of secondary metabolites, biosynthesis of amino acids, carbon metabolism and biosynthesis of cofactors for the models using k_{cat} and k_{max}^{vivo} values (Figure S2). The GO term biosynthesis of secondary metabolites encompasses enzymes catalyzing a broad set of reactions, that may explain its significant enrichment in all tested protein sets. The five proteins with the highest relative error (Table S4) showed overlap between the two model parameterizations and were part of purine metabolism, hexose metabolism, and porphyrin-containing compound metabolism. These comparisons demonstrated that the protein reserve ratios did not exhibit distinct patterns based on which metabolic pathway(s) in which the proteins are involved.

To further compare the two groups of proteins, we examined whether the differences in the relative error of predictions were associated with the coefficient of variation (CV) of protein reserve ratios calculated from the models using either k_{cat} and k_{max}^{vivo} values. For all compared scenarios, many proteins with low CV also had high relative errors. However, we did not identify a statistically significant difference in CV between proteins with low and high relative errors in all scenarios except the *E. coli* model using k_{max}^{vivo} values (Pairwise Wilcoxon rank sum test with Bonferroni correction, Figure S3). Taken together, these results indicated that proteins with low or high prediction errors are used in similar metabolic contexts, suggesting that other physiological factors might contribute to the predictive performance of their abundances.

CAMEL accurately predicts in vivo protein abundances

Next, we used the predicted distributions of protein reserve ratios, $\hat{\boldsymbol{\phi}}$, to estimate protein abundance, E'_s, which is given by $E'_{s,i} = \frac{E_{s,i}}{\hat{\phi}_{s,i}}$, where $E_{s,i}$ is the condition-specific protein abundance predicted from the pcGEM. The obtained protein abundance estimates were compared to experimental measurements (Chen and Nielsen, 2021; Davidi et al., 2016) by calculating the Pearson correlation coefficient of $\log_{10}$-transformed protein abundances for each growth condition separately (shown in Tables S5-S8) as well as the average across all conditions (Figure 4).

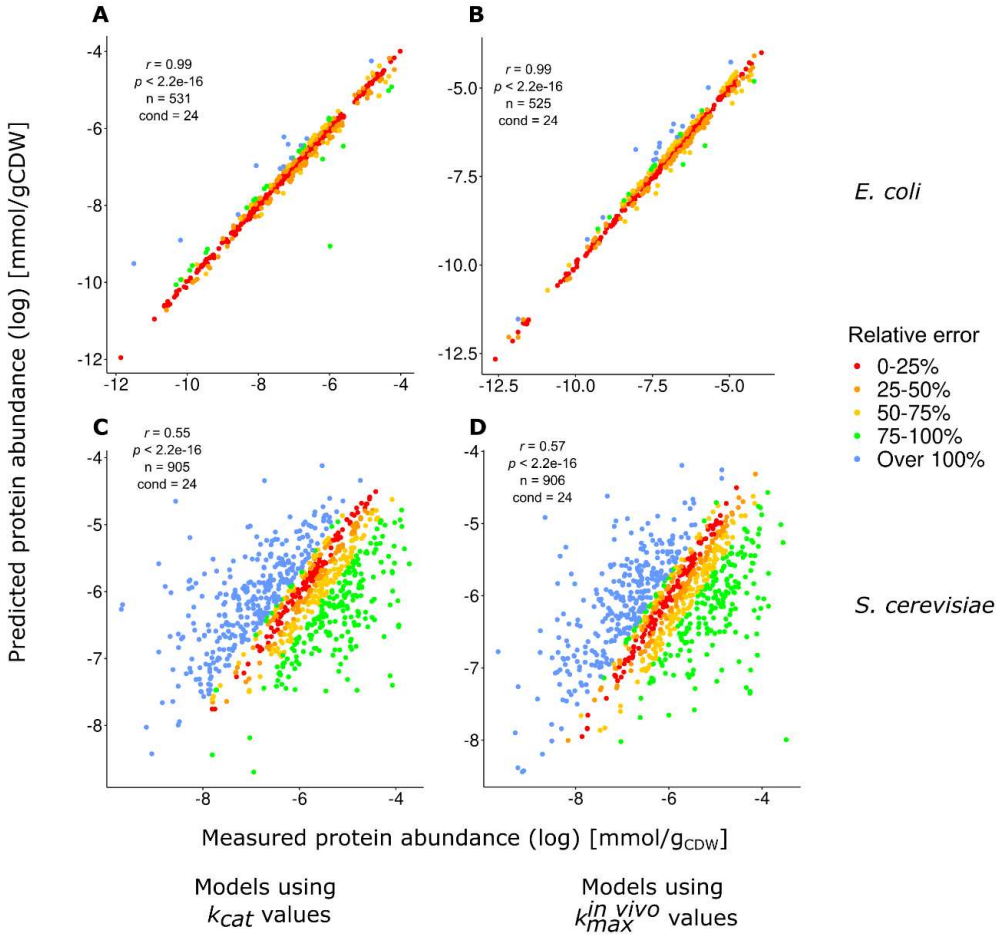

Figure 4. Comparison of measured and predicted protein abundances. The colour scheme indicates the relative error when comparing the measured (E_s^{exp}) and predicted protein abundances ($E_s^{exp\prime}$). Relative error was calculated using protein abundances in nominal scale. Panel (A) shows predictions using the *E. coli* model with *in vitro* k_{cat} values (A), and in vivo k_{max}^{vivo} values (B). The panels (C) and (D), depict the prediction performance of the *S. cerevisiae* model with k_{cat} values (C) and k_{max}^{vivo} values (D).

For *E. coli*, we found a high and significant average value for the Pearson correlation coefficient of 0.987 (p-values < 2.5e-5) over the considered conditions using the models for $\hat{\boldsymbol{\phi}}$ based on the model parameterized with k_{cat} values, and 0.994 (p-values < 3.1e-11) using the model parameterized with k_{max}^{vivo} values. In *S. cerevisiae*, using the k_{cat} parameterization, we found a smaller, yet significant average Pearson correlation of 0.483 (p-values < 0.05 for all significant correlations) over the considered conditions; using the pcGEM with k_{max}^{vivo} values resulted in an average Pearson's correlation of 0.513 (p-values < 0.03 for all significant correlations).

Here, too, we took a closer look at proteins with very high relative errors and how their CV compare to proteins with very low relative errors. In *E. coli*, the proteins with the highest relative error showed values larger than 1000% (Table S9). These proteins were part of the chorismate biosynthesis, fatty acid beta-oxidation, pentose phosphate pathway and peptidoglycan biosynthesis when predictions were based on the pcGEM with k_{cat} values. In the case when k_{max}^{vivo} parameterization was employed, these proteins were part of fatty acid biosynthesis, pentose phosphate pathway and purine metabolism. For *S. cerevisiae*, the proteins with the highest errors showed more extreme values than in the case of *E. coli* (Table S10). For both pcGEM parameterizations, the proteins with high error are related to amino acid biosynthesis. As also seen by Lu et al., (2019) and Xia et al., (2022), this observation could be attributed to how different strains and growth conditions affect amino acid biosynthesis pathways. Lu et al. (2019) details that differences in energy pathways used for ATP regeneration directly impacts the biosynthesis of amino acids, given that central metabolic pathways are the main supplier of precursor molecules. Indeed, we observed that some of the proteins with the highest relative error are also related to energy metabolism, such as the pyruvate decarboxylase isozyme 1 and the pyruvate dehydrogenase complex protein X component (Table S10). Xia et al. (2022) have also made a similar observation, stating that the proteome fraction allocated to amino acid biosynthesis is metabolism-dependent. As the cell increases protein translation, the proteome fraction allocated to other sections of metabolism is decreased. However, for the growth conditions we have chosen, the growth rates did not exceed the threshold to trigger the onset of the Crabtree effect. Further, the models were not constrained on

amino acid exchange reactions, allowing the model to freely uptake required amino acids. This could explain the high error on these proteins, since they are not being actively used for amino acid biosynthesis. Regarding the CV, we observed no significant difference between proteins with very high or very low errors (Figure S4) (pairwise Wilcoxon rank sum test with Bonferroni correction), except for the yeast pcGEM with k_{max}^{vivo} values (p-value < 0.027).

The FVA analysis revealed that for *S. cerevisiae*, there is little variation on metabolic flux through central metabolic pathways (median flux ratio of 16.3 across conditions). For *E. coli*, there was higher variability (median flux ratio of 64.9 across conditions), likely due to the constraints used, since less exchange reactions of central metabolites were constrained than for *S. cerevisiae*. For the enzyme usage pseudo-reactions, we assessed if enzymes with high relative error could be related to pseudo-reactions with high variability. For both *E. coli* and *S. cerevisiae*, using either k_{cat} or k_{max}^{vivo} values, there was no relation between the enzyme usage variability and the relative error. Proteins with either high or low relative errors exhibited similar variability. Some proteins with high variability displayed low relative errors, although the difference in enzyme usage variability between the proteins with low variability and proteins with high variability was very small (median ratio variability of 1.0) (Table S11).

To further demonstrate the capabilities of CAMEL, we validated its predictive performance using data from unseen conditions. We predicted the E_s and ϕ of mutant *E. coli* strains subjected to ALE. We then used the predictions to calculate the $E_s^{exp\prime}$ and compared with experimental values using the Pearson correlation coefficient. For the ML trained with the k_{cat}-parameterized pcGEM data, we obtained a value of 0.626 for the Pearson correlation coefficient, while for the ML trained with the k_{max}^{vivo}-parameterized pcGEM data, we obtained a value of 0.533 for the Pearson correlation coefficient. This result shows that CAMEL is able to obtain good predictions for genetically modified strains, where the rationale behind metabolic optimality may not hold.

Previous attempts for predicting protein abundance using constraint-based models achieved good predictive performances, but they underestimated *in vivo* protein abundance. The predictions performed by Adadi et al. (2012) using the MOMENT approach achieved a Pearson's correlation of 0.84 between predicted protein contents and gene expression data of *E. coli* grown on a glucose minimal

medium. We showed that CAMEL with k_{max}^{vivo} parameterized pcGEM resulted in Pearson correlations of over 0.9 for *E. coli* over different conditions, demonstrating considerably improved performance by using the protein reserve ratios to predict protein usage values compatible to the *in vivo* protein usage. In addition, CAMEL does not require the use of any additional molecular read-outs, thus saving valuable resources for making predictions in unseen conditions. We note that the performance of CAMEL is comparable to that of Resource Balance Analysis (RBA) model of *Bacillus subtilis*, achieving an R^2 value of 0.94 (Goelzer et al., 2015); however, predictions from RBA models consider more information about transcription, protein translation, translocation, compartmentalization, folding and thermostability, which are difficult to parameterize even in prokaryotic model organisms. CAMEL, on the other hand, depends on pcGEMs, quantitative proteomics data and coding sequences for generating its predictions.

Attempts to predict protein abundance using machine learning have previously achieved good predictive performances but rely on either transcriptomics data or sequence-derived features, which have low correlation to protein content, and are invariable to changes in environmental conditions, respectively (Ferreira et al., 2021; Li et al., 2019). For instance, the model constructed by Terai and Asai (Terai and Asai, 2020), based on three different algorithms, was trained to predict protein abundance in *E. coli* using sequence-derived features and mRNA structure information. The model performance was assessed by Spearman correlation, which ranged from 0.55 to 0.71. Predictions for *S. cerevisiae* were carried out by the Bayesian network constructed by Mehdi et al. (Mehdi et al., 2014), which combined transcriptomics measurements and sequence-derived features, using data from *S. cerevisiae* and *Schizosaccharomyces pombe*. The predictions were evaluated by Spearman correlation, which ranged from 0.61 to 0.77. Ferreira et al. (Ferreira et al., 2021) generated a predictive model using only codon usage metrics as features, which achieved a Spearman correlation of 0.74. Although the resulting correlations were higher than those from CAMEL, both studies relied only on data obtained from optimal growth conditions. However, CAMEL can be effectively used to make predictions of protein abundance in sub-optimal scenarios, which is a notoriously difficult endeavor (Li et al., 2019). Given that predicted protein abundance based on the protein reserve ratios are in good agreement with

experimental values, these results indicated that the coupling of pcGEMs and machine learning can accurately predict the protein allocation *in vivo*.

By coupling constraint-based approaches and machine learning approaches, we demonstrated that our proposed approach, CAMEL, generated accurate predictions of the protein reserve ratios. In addition, we showed that in combination with constraint-based models, CAMEL leverages these machine learning models to obtain a prediction of protein abundance that overcome notable limitations inherent to pcGEMs (e.g., over-accumulation of enzymes, uncertainty in k_{cat} values). As its most notable merit, CAMEL predicts protein abundance by relying on physiological data, used in the prediction of fluxes, rather than requiring additional omics measurements such as gene expression. Therefore, CAMEL is readily applicable to any condition given that measurements of growth rates and exchange fluxes are available, widening the possibilities for its biotechnological applications (e.g., design of strains). In addition, as our findings demonstrated, CAMEL results in good predictions for protein abundance for sub-optimal conditions, opening the possibility for using these predictions as features to model other complex physiological traits.

Code and data availability

The code and data can be found in the GitHub repository: https://github.com/mauricioamf/CAMEL

Acknowledgments

This study was financed in part by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior – Brasil (CAPES) – Finance Code 001. P.W. and Z.N. acknowledge the support from the Research Focus Area “Evolutionary Systems Biology” of the University of Potsdam. M.A. and Z.N. acknowledge the support of the Max Planck Society. M.A. and Z.N were supported by the European Union’s Horizon 2020 research and innovation programme, project PlantaSYST (SGA-CSA No. 739582 under FPA No. 664620).

References

Adadi, R., Volkmer, B., Milo, R., Heinemann, M., Shlomi, T., 2012. Prediction of Microbial Growth Rate versus Biomass Yield by a Metabolic Network with Kinetic

Parameters. PLoS Comput Biol 8, e1002575. https://doi.org/10.1371/JOURNAL.PCBI.1002575

Alter, T.B., Blank, L.M., Ebert, B.E., 2021. Proteome Regulation Patterns Determine Escherichia coli Wild-Type and Mutant Phenotypes. mSystems 6. https://doi.org/10.1128/msystems.00625-20

Ashburner, M., Ball, C.A., Blake, J.A., Botstein, D., Butler, H., Cherry, J.M., Davis, A.P., Dolinski, K., Dwight, S.S., Eppig, J.T., Harris, M.A., Hill, D.P., Issel-Tarver, L., Kasarskis, A., Lewis, S., Matese, J.C., Richardson, J.E., Ringwald, M., Rubin, G.M., Sherlock, G., 2000. Gene Ontology: tool for the unification of biology. Nat Genet 25, 25. https://doi.org/10.1038/75556

Beg, Q.K., Vazquez, A., Ernst, J., De Menezes, M.A., Bar-Joseph, Z., Barabási, A.L., Oltvai, Z.N., 2007. 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. https://doi.org/10.1073/PNAS.0609845104/SUPPL_FILE/09845DATASET5.XLS

Bekiaris, P.S., Klamt, S., 2020. Automatic construction of metabolic models with enzyme constraints. BMC Bioinformatics 21, 1–13. https://doi.org/10.1186/S12859-019-3329-9/TABLES/2

Bernstein, D.B., Akkas, B., Price, M.N., Arkin, A.P., 2023. Critical assessment of E. coli genome-scale metabolic model with high-throughput mutant fitness data. bioRxiv 2023.01.05.522875. https://doi.org/10.1101/2023.01.05.522875

Bruggeman, F.J., Planqué, R., Molenaar, D., Teusink, B., 2020. Searching for principles of microbial physiology. FEMS Microbiol Rev 44, 821–844. https://doi.org/10.1093/FEMSRE/FUAA034

Calderón-Celis, F., Encinar, J.R., Sanz-Medel, A., 2018. Standardization approaches in absolute quantitative proteomics with mass spectrometry. Mass Spectrom Rev 37, 715–737. https://doi.org/10.1002/mas.21542

Carbon, S., Douglass, E., Good, B.M., Unni, D.R., Harris, N.L., Mungall, C.J., Basu, S., Chisholm, R.L., Dodson, R.J., Hartline, E., Fey, P., Thomas, P.D., Albou, L.P., Ebert, D., Kesling, M.J., Mi, H., Muruganujan, A., Huang, X., Mushayahama, T., LaBonte, S.A., Siegele, D.A., Antonazzo, G., Attrill, H., Brown, N.H., Garapati, P., Marygold, S.J., Trovisco, V., dos Santos, G., Falls, K., Tabone, C., Zhou, P., Goodman, J.L., Strelets, V.B., Thurmond, J., Garmiri, P., Ishtiaq, R., Rodríguez-López,

M., Acencio, M.L., Kuiper, M., Lægreid, A., Logie, C., Lovering, R.C., Kramarz, B., Saverimuttu, S.C.C., Pinheiro, S.M., Gunn, H., Su, R., Thurlow, K.E., Chibucos, M., Giglio, M., Nadendla, S., Munro, J., Jackson, R., Duesbury, M.J., Del-Toro, N., Meldal, B.H.M., Paneerselvam, K., Perfetto, L., Porras, P., Orchard, S., Shrivastava, A., Chang, H.Y., Finn, R.D., Mitchell, A.L., Rawlings, N.D., Richardson, L., Sangrador-Vegas, A., Blake, J.A., Christie, K.R., Dolan, M.E., Drabkin, H.J., Hill, D.P., Ni, L., Sitnikov, D.M., Harris, M.A., Oliver, S.G., Rutherford, K., Wood, V., Hayles, J., Bähler, J., Bolton, E.R., de Pons, J.L., Dwinell, M.R., Hayman, G.T., Kaldunski, M.L., Kwitek, A.E., Laulederkind, S.J.F., Plasterer, C., Tutaj, M.A., Vedi, M., Wang, S.J., D'Eustachio, P., Matthews, L., Balhoff, J.P., Aleksander, S.A., Alexander, M.J., Cherry, J.M., Engel, S.R., Gondwe, F., Karra, K., Miyasato, S.R., Nash, R.S., Simison, M., Skrzypek, M.S., Weng, S., Wong, E.D., Feuermann, M., Gaudet, P., Morgat, A., Bakker, E., Berardini, T.Z., Reiser, L., Subramaniam, S., Huala, E., Arighi, C.N., Auchincloss, A., Axelsen, K., Argoud-Puy, G., Bateman, A., Blatter, M.C., Boutet, E., Bowler, E., Breuza, L., Bridge, A., Britto, R., Bye-A-Jee, H., Casas, C.C., Coudert, E., Denny, P., Es-Treicher, A., Famiglietti, M.L., Georghiou, G., Gos, A.N., Gruaz-Gumowski, N., Hatton-Ellis, E., Hulo, C., Ignatchenko, A., Jungo, F., Laiho, K., Le Mercier, P., Lieberherr, D., Lock, A., Lussi, Y., MacDougall, A., Ma-Grane, M., Martin, M.J., Masson, P., Natale, D.A., Hyka-Nouspikel, N., Orchard, S., Pedruzzi, I., Pourcel, L., Poux, S., Pundir, S., Rivoire, C., Speretta, E., Sundaram, S., Tyagi, N., Warner, K., Zaru, R., Wu, C.H., Diehl, A.D., Chan, J.N., Grove, C., Lee, R.Y.N., Muller, H.M., Raciti, D., van Auken, K., Sternberg, P.W., Berriman, M., Paulini, M., Howe, K., Gao, S., Wright, A., Stein, L., Howe, D.G., Toro, S., Westerfield, M., Jaiswal, P., Cooper, L., Elser, J., 2021. The Gene Ontology resource: enriching a GOld mine. Nucleic Acids Res 49, D325–D334. https://doi.org/10.1093/NAR/GKAA1113

Chang, A., Jeske, L., Ulbrich, S., Hofmann, J., Koblitz, J., Schomburg, I., Neumann-Schaal, M., Jahn, D., Schomburg, D., 2021. BRENDA, the ELIXIR core data resource in 2021: New developments and updates. Nucleic Acids Res 49, D498–D508. https://doi.org/10.1093/nar/gkaa1025

Chen, T., Guestrin, C., 2016. XGBoost: A scalable tree boosting system. Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining 13-17-Augu, 785–794. https://doi.org/10.1145/2939672.2939785

Chen, Y., Nielsen, J., 2021. In vitro turnover numbers do not reflect in vivo activities of yeast enzymes. Proc Natl Acad Sci U S A 118, e2108391118. https://doi.org/10.1073/PNAS.2108391118/SUPPL_FILE/PNAS.2108391118.SD08.XLSX

Davidi, D., Noor, E., Liebermeister, W., Bar-Even, A., Flamholz, A., Tummler, K., Barenholz, U., Goldenfeld, M., Shlomi, T., Milo, R., 2016. Global characterization of in vivo enzyme catalytic rates and their correspondence to in vitro kcat measurements. Proc Natl Acad Sci U S A 113, 3401–3406. https://doi.org/10.1073/pnas.1514240113

Di Bartolomeo, F., Malina, C., Campbell, K., Mormino, M., Fuchs, J., Vorontsov, E., Gustafsson, C.M., Nielsen, J., 2020. Absolute yeast mitochondrial proteome quantification reveals trade-off between biosynthesis and energy generation during diauxic shift. Proc Natl Acad Sci U S A 117, 7524–7535. https://doi.org/10.1073/PNAS.1918216117/SUPPL_FILE/PNAS.1918216117.SD07.XLSX

Domenzain, I., Sánchez, B., Anton, M., Kerkhoven, E.J., Millán-Oropeza, A., Henry, C., Siewers, V., Morrissey, J.P., Sonnenschein, N., Nielsen, J., 2022. Reconstruction of a catalogue of genome-scale metabolic models with enzymatic constraints using GECKO 2.0. Nat Commun 13, 1–13. https://doi.org/10.1038/s41467-022-31421-1

Eraslan, B., Wang, D., Gusic, M., Prokisch, H., Orn, B., Hallstr€ Om, M., Uhl En, M., Asplund, A., Pont En, F., Wieland, T., Hopf, T., Hahne, H., Kuster, B., Gagneur, J., 2019. Quantification and discovery of sequence determinants of protein-per-mRNA amount in 29 human tissues. Mol Syst Biol 15, e8513. https://doi.org/10.15252/MSB.20188513

Ferreira, M., Ventorim, R., Almeida, E., Silveira, S., Silveira, W., 2021. Protein Abundance Prediction Through Machine Learning Methods. J Mol Biol 433, 167267. https://doi.org/10.1016/J.JMB.2021.167267

Goelzer, A., Muntel, J., Chubukov, V., Jules, M., Prestel, E., Nölker, R., Mariadassou, M., Aymerich, S., Hecker, M., Noirot, P., Becher, D., Fromion, V., 2015. Quantitative prediction of genome-wide resource allocation in bacteria. Metab Eng 32, 232–243. https://doi.org/10.1016/J.YMBEN.2015.10.003

Heckmann, D., Campeau, A., Lloyd, C.J., Phaneuf, P. V., Hefner, Y., Carrillo-Terrazas, M., Feist, A.M., Gonzalez, D.J., Palsson, B.O., 2020. Kinetic profiling of metabolic specialists demonstrates stability and consistency of in vivo enzyme turnover numbers. Proc Natl Acad Sci U S A 117, 23182–23190. https://doi.org/10.1073/PNAS.2001562117/SUPPL_FILE/PNAS.2001562117.SD01.XLSX

Heckmann, D., Lloyd, C.J., Mih, N., Ha, Y., Zielinski, D.C., Haiman, Z.B., Desouki, A.A., Lercher, M.J., Palsson, B.O., 2018. Machine learning applied to enzyme turnover numbers reveals protein structural correlates and improves metabolic models. Nat Commun 9, 1–10. https://doi.org/10.1038/s41467-018-07652-6

Kintaka, R., Makanae, K., Namba, S., Kato, H., Kito, K., Ohnuki, S., Ohya, Y., Andrews, B.J., Boone, C., Moriya, H., 2020. Genetic profiling of protein burden and nuclear export overload. Elife 9, 1–22. https://doi.org/10.7554/ELIFE.54080

Kültz, D., 2020. Evolution of cellular stress response mechanisms. J Exp Zool A Ecol Integr Physiol 333, 359–378. https://doi.org/10.1002/JEZ.2347

Lahtvee, P.J., Sánchez, B.J., Smialowska, A., Kasvandik, S., Elsemman, I.E., Gatto, F., Nielsen, J., 2017. Absolute Quantification of Protein and mRNA Abundances Demonstrate Variability in Gene-Specific Translation Efficiency in Yeast. Cell Syst 4, 495-504.e5. https://doi.org/10.1016/J.CELS.2017.03.003

Le, T.T., Fu, W., Moore, J.H., 2020. Scaling tree-based automated machine learning to biomedical big data with a feature set selector. Bioinformatics 36, 250–256. https://doi.org/10.1093/BIOINFORMATICS/BTZ470

Li, G., Hu, Y., Jan Zrimec, Luo, H., Wang, H., Zelezniak, A., Ji, B., Nielsen, J., 2021. Bayesian genome scale modelling identifies thermal determinants of yeast metabolism. Nat Commun 12, 1–12. https://doi.org/10.1038/s41467-020-20338-2

Li, H., Siddiqui, O., Zhang, H., Guan, Y., 2019. Joint learning improves protein abundance prediction in cancers. BMC Biol 17, 1–14. https://doi.org/10.1186/S12915-019-0730-9/FIGURES/6

Lill, J.R., Mathews, W.R., Rose, C.M., Schirle, M., 2021. Proteomics in the pharmaceutical and biotechnology industry: a look to the next decade. https://doi.org/10.1080/14789450.2021.1962300 18, 503–526. https://doi.org/10.1080/14789450.2021.1962300

Liu, Y., Lu, S., Liu, K., Wang, S., Huang, L., Guo, L., 2019. Proteomics: a powerful tool to study plant responses to biotic stress. Plant Methods 2019 15:1 15, 1–20. https://doi.org/10.1186/S13007-019-0515-8

Lu, H., Li, F., Sánchez, B.J., Zhu, Z., Li, G., Domenzain, I., Marcišauskas, S., Anton, P.M., Lappa, D., Lieven, C., Beber, M.E., Sonnenschein, N., Kerkhoven, E.J., Nielsen, J., 2019. A consensus S. cerevisiae metabolic model Yeast8 and its ecosystem for comprehensively probing cellular metabolism. Nat Commun 10. https://doi.org/10.1038/s41467-019-11581-3

McCloskey, D., Xu, S., Sandberg, T.E., Brunk, E., Hefner, Y., Szubin, R., Feist, A.M., Palsson, B.O., 2018a. Adaptive laboratory evolution resolves energy depletion to maintain high aromatic metabolite phenotypes in Escherichia coli strains lacking the Phosphotransferase System. Metab Eng 48, 233–242. https://doi.org/10.1016/J.YMBEN.2018.06.005

McCloskey, D., Xu, S., Sandberg, T.E., Brunk, E., Hefner, Y., Szubin, R., Feist, A.M., Palsson, B.O., 2018b. Adaptation to the coupling of glycolysis to toxic methylglyoxal production in tpiA deletion strains of Escherichia coli requires synchronized and counterintuitive genetic changes. Metab Eng 48, 82–93. https://doi.org/10.1016/J.YMBEN.2018.05.012

McCloskey, D., Xu, S., Sandberg, T.E., Brunk, E., Hefner, Y., Szubin, R., Feist, A.M., Palsson, B.O., 2018c. Multiple optimal phenotypes overcome redox and glycolytic intermediate metabolite imbalances in Escherichia coli pgi knockout evolutions. Appl Environ Microbiol 84, 823–841. https://doi.org/10.1128/AEM.00823-18/SUPPL_FILE/ZAM019188775SD8.CSV

McCloskey, D., Xu, S., Sandberg, T.E., Brunk, E., Hefner, Y., Szubin, R., Feist, A.M., Palsson, B.O., 2018d. Growth adaptation of gnd and sdhCB Escherichia coli deletion strains diverges from a similar initial perturbation of the transcriptome. Front Microbiol 9, 398699. https://doi.org/10.3389/FMICB.2018.01793/BIBTEX

Mehdi, A.M., Patrick, R., Bailey, T.L., Boden, M., 2014. Predicting the Dynamics of Protein Abundance. Molecular & Cellular Proteomics 13, 1330–1340. https://doi.org/10.1074/MCP.M113.033076

Mergner, J., Frejno, M., List, M., Papacek, M., Chen, X., Chaudhary, A., Samaras, P., Richter, S., Shikata, H., Messerer, M., Lang, D., Altmann, S., Cyprys, P., Zolg, D.P., Mathieson, T., Bantscheff, M., Hazarika, R.R., Schmidt, T., Dawid, C.,

Dunkel, A., Hofmann, T., Sprunck, S., Falter-Braun, P., Johannes, F., Mayer, K.F.X., Jürgens, G., Wilhelm, M., Baumbach, J., Grill, E., Schneitz, K., Schwechheimer, C., Kuster, B., 2020. Mass-spectrometry-based draft of the Arabidopsis proteome. Nature 579, 409–414. https://doi.org/10.1038/s41586-020-2094-2

Nielsen, J., 2019. Yeast Systems Biology: Model Organism and Cell Factory. Biotechnol J 14, 1–9. https://doi.org/10.1002/biot.201800421

Novoa, E.M., Jungreis, I., Jaillon, O., Kellis, M., Leitner, T., 2019. Elucidation of Codon Usage Signatures across the Domains of Life. Mol Biol Evol 36, 2328–2339. https://doi.org/10.1093/molbev/msz124

Otto, A., Becher, D., Schmidt, F., 2014. Quantitative proteomics in the field of microbiology. Proteomics 14, 547–565. https://doi.org/10.1002/pmic.201300403

Pappireddi, N., Martin, L., Wühr, M., 2019. A Review on Quantitative Multiplexed Proteomics. ChemBioChem 20, 1210–1224. https://doi.org/10.1002/cbic.201800650

Peebo, K., Valgepea, K., Maser, A., Nahku, R., Adamberg, K., Vilu, R., 2015. Proteome reallocation in Escherichia coli with increasing specific growth rate. Mol Biosyst 11, 1184–1193. https://doi.org/10.1039/C4MB00721B

Sánchez, B.J., Zhang, C., Nilsson, A., Lahtvee, P., Kerkhoven, E.J., Nielsen, J., 2017. Improving the phenotype predictions of a yeast genome-scale metabolic model by incorporating enzymatic constraints. Mol Syst Biol 13, 935. https://doi.org/10.15252/msb.20167411

Schmidt, A., Kochanowski, K., Vedelaar, S., Ahrné, E., Volkmer, B., Callipo, L., Knoops, K., Bauer, M., Aebersold, R., Heinemann, M., 2016. The quantitative and condition-dependent Escherichia coli proteome. Nat Biotechnol 34, 104–110. https://doi.org/10.1038/nbt.3418

Schubert, O.T., Röst, H.L., Collins, B.C., Rosenberger, G., Aebersold, R., 2017. Quantitative proteomics: challenges and opportunities in basic and applied research. Nature Protocols 2017 12:7 12, 1289–1294. https://doi.org/10.1038/nprot.2017.040

Terai, G., Asai, K., 2020. Improving the prediction accuracy of protein abundance in Escherichia coli using mRNA accessibility. Nucleic Acids Res 48, e81–e81. https://doi.org/10.1093/nar/gkaa481

Torres-García, W., Zhang, W., Runger, G.C., Johnson, R.H., Meldrum, D.R., 2009. Integrative analysis of transcriptomic and proteomic data of Desulfovibrio vulgaris:

a non-linear model to predict abundance of undetected proteins. Bioinformatics 25, 1905–1914. https://doi.org/10.1093/BIOINFORMATICS/BTP325

Valgepea, K., Adamberg, K., Seiman, A., Vilu, R., 2013. Escherichia coli achieves faster growth by increasing catalytic and translation rates of proteins. Mol Biosyst 9, 2344–2358. https://doi.org/10.1039/C3MB70119K

Wendering, P., Arend, M., Razaghi-Moghadam, Z., Nikoloski, Z., 2023. Data integration across conditions improves turnover number estimates and metabolic predictions. Nature Communications 2023 14:1 14, 1–12. https://doi.org/10.1038/s41467-023-37151-2

Wu, T., Hu, E., Xu, S., Chen, M., Guo, P., Dai, Z., Feng, T., Zhou, L., Tang, W., Zhan, L., Fu, X., Liu, S., Bo, X., Yu, G., 2021. clusterProfiler 4.0: A universal enrichment tool for interpreting omics data. Innovation 2, 100141. https://doi.org/10.1016/j.xinn.2021.100141

Xia, J., Sánchez, B.J., Chen, Y., Campbell, K., Kasvandik, S., Nielsen, J., 2022. Proteome allocations change linearly with the specific growth rate of Saccharomyces cerevisiae under glucose limitation. Nature Communications 2022 13:1 13, 1–12. https://doi.org/10.1038/s41467-022-30513-2

Xu, R., Razaghi-Moghadam, Z., Nikoloski, Z., 2021. Maximization of non-idle enzymes improves the coverage of the estimated maximal in vivo enzyme catalytic rates in Escherichia coli. Bioinformatics 37, 3848–3855. https://doi.org/10.1093/BIOINFORMATICS/BTAB575

Yu, G., Wang, L.G., Han, Y., He, Q.Y., 2012. ClusterProfiler: An R package for comparing biological themes among gene clusters. OMICS 16, 284–287. https://doi.org/10.1089/OMI.2011.0118/ASSET/IMAGES/LARGE/FIGURE1.JPEG

Yu, R., Campbell, K., Pereira, R., Björkeroth, J., Qi, Q., Vorontsov, E., Sihlbom, C., Nielsen, J., 2020. Nitrogen limitation reveals large reserves in metabolic and translational capacities of yeast. Nat Commun 11, 1–12. https://doi.org/10.1038/s41467-020-15749-0

3.4 Integrating promiscuous enzyme activity and underground metabolism in protein-constrained models

Original text not yet submitted to a journal or preprint repository.

Abstract

The integration of enzyme parameters in constraint-based models have significantly enhanced their predictive capabilities, facilitating the prediction of enzyme resource usage and distribution. However, current approaches largely neglect the possible set of promiscuous enzyme activities that jointly comprise the so-called underground metabolism. To allow enzyme-constrained study of underground metabolism, we developed the CORAL Toolbox that includes enzyme subpools for each reaction, increasing the resolution in which enzyme resource allocation can be modelled. Applying CORAL with a genome-scale metabolic model of *Escherichia coli*, we found that underground metabolism results in larger flexibility in metabolic fluxes and enzyme usage compared to classical models. Next, we investigated how enzyme resources are distributed to side reactions when the main reaction is knocked out using the model and found that the pools of most side reactions, with few exceptions, show a slight change in value. We also investigated how knocking out pairs of main reactions affects growth, and found that enzymes without promiscuous activity have more impact on growth than enzymes with promiscuous activity, suggesting that promiscuous enzyme activity is vital to maintain robust metabolic function and growth.

Introduction

Enzymes are the workhorses of metabolism. They catalyse the conversion of substrates into products for the majority of metabolic reactions. Although some enzymes

are highly specific with respect to reactions they catalyse, there is notable enzyme promiscuity, with estimates of over thousands of promiscuous activities (COPLEY; NEWTON; WIDNEY, 2023). Promiscuous enzymes catalyse more than a single reaction by binding with smaller affinity to other substrates (TAWFIK, 2010). As a consequence of lower substrate affinity, the side reactions catalysed by promiscuous enzymes occur at a lower rate than the main reaction. Promiscuous enzymes also exhibit lower catalytic efficiency for the side in comparison to the main reaction (COPLEY, 2017). As a result, it has been taught that side reactions are mostly physiologically irrelevant (COPLEY, 2015). However, these side reactions still happen often enough to form an alternative metabolic network termed underground metabolism (D'ARI; CASADESÚS, 1998). This underground metabolic network serves an important role in evolution, providing the reservoir of enzyme functions to evolve via natural selection (NOTEBAART et al., 2014). Evidence indicates that gene duplication events contribute to underground metabolism by creating copies of the enzyme, with one copy evolving higher affinity to the substrate of a certain side reaction (GLASNER; TRUONG; MORSE, 2020). Further, underground metabolism can also be used for biotechnological purposes, aiding laboratory adaptive evolution (ALE) experiments (GUZMÁN et al., 2019) and guiding metabolic engineering efforts (KOVÁCS et al., 2022).

Constraint-based approaches rely on optimization principles to predict and study metabolic phenotypes using genome-scale metabolic models (GEMs) (ORTH; THIELE; PALSSON, 2010). GEM-based investigations of underground metabolism have revealed the connectivity between native and underground metabolism and how underground reactions contributes to adaptation to new environments (NOTEBAART et al., 2014), improved gap-filling of GEMs (PAN; REED, 2018), and to the design of metabolic engineering strategies (KOVÁCS et al., 2022). While insightful, these

investigations were performed using conventional GEMs, which do not consider constraints on the available enzyme abundance and enzyme catalytic rates. Consideration of these constraints has resulted in the generation of protein-constrained GEMs (pcGEMs) that have been shown to improve predictive performance (ADADI et al., 2012; SÁNCHEZ et al., 2017; WENDERING; NIKOLOSKI, 2022). However, it remains elusive how the metabolic network allocates enzyme resources between main and side reactions.

The GECKO 3 (CHEN et al., 2024) formulation of pcGEMs represents each enzyme as a pseudometabolite that participates in a reaction, with a stoichiometric coefficient given by the ratio between the molecular weight of the enzyme (MW) and its turnover number, k_{cat}, for the reaction. For each enzyme, a pseudoreaction draws from the total protein pool an amount corresponding to its usage in the model. The total protein pool is then obtained from an exchange reaction. In this formulation, an enzyme E1 that catalyses two or more reactions uses the same E1 pseudometabolite in each reaction, indicating that the same amount of enzyme is used for all reactions catalysed by the enzyme. However, since most enzymes are already occupied by their main substrates, the availability of enzyme resources to side reactions may be drastically reduced. This issue is also not accounted in the existing pcGEM approaches.

To address these questions, here we propose an approach for modelling promiscuous enzyme activity in pcGEMs, termed **CO**nstraint-based p**R**omiscuous enzyme **A**nd underground metabolism mode**L**ling (CORAL) Toolbox. The CORAL Toolbox builds on GECKO 3 to predict enzyme allocation while ensuring that a subpool allocation to a promiscuous enzyme is used for a particular reaction. We show that CORAL can predict the distribution of resources among main and side reactions,

making it a useful approach to understand promiscuous enzyme activity and underground metabolism.

Material and methods

Refining the iML1515 model and reconstructing the pcGEM

To reconstruct a GEM accounting for underground metabolism, we manually curated the *Escherichia coli* GEM iML1515 (MONK et al., 2017) to include underground reactions. We included the reactions from the underground iJO1366 (ORTH et al., 2011) model created by Kovács et al. (2022). These were predicted by the PROPER algorithm, with 20% of these reactions validated experimentally (NOTEBAART et al., 2014; OBERHARDT et al., 2016). We integrated these reactions, matched all annotations from iJO1366 to the format used by iML1515, and standardized the gene-protein-reaction (GPR) rules, resulting in the iML1515u model. These reactions account for the production of 160 different value-added chemicals, such as polylactic acid, 3-hydroxypropanoate and 3-hydroxybutryolactone, used for bioplastics production; and hydroquinone, (R)-3-hydroxybutanoate and 4-hydroxy-3-methoxy-benzaldehyde, used in the pharmaceutical industry.

Next, we reconstructed the protein-constrained version of iML1515u using the GECKO Toolbox 3 (CHEN et al., 2024). We populated the adapter file with *E. coli*-specific information, using a maximum growth rate when not constrained by nutrient uptake of 0.6 h^{-1} (SCHMIDT et al., 2016), total protein content (P_{tot}) of 0.61 $g_{protein}/g_{DW}$ (VALGEPEA et al., 2013), and the default value of 0.5 for average enzyme saturation factor (σ) and fraction of enzymes in the model (f). We followed the steps described to reconstruct the full ecModel, which has constraints on individual enzymes. For all

reactions in the model, we integrated k_{cat} values predicted by DLKcat (LI et al., 2022). We followed the protocol until Stage 2 since Stage 3 deal with tuning model parameters and we wanted to preserve the original DLKcat-predicted k_{cat} values, and Stage 4 adjust growth parameters using quantitative proteomics data. All model refinements were performed using the COBRA Toolbox 3 (HEIRENDT et al., 2019) and the RAVEN Toolbox 2 (WANG et al., 2018) in MATLAB (The MathWorks Inc., Natick, Massachusetts).

Restructuring the pcGEM to account for enzyme usage in underground and promiscuous reactions

To account for the enzyme usage for individual reactions catalysed by a promiscuous enzyme, we developed the CORAL Toolbox. CORAL modifies how enzymes are used by splitting the pool of an enzyme that catalyses more than one reaction into multiple subpools, with each subpool being responsible for the enzyme resources of only one reaction. The restructuring of a GECKO3-constructed pcGEM is done in three steps: (i) simplifying GPR rules; (ii) splitting enzyme pools into subpools for all reactions; and (iii) updating enzyme information on the model.

The first step of the restructuring entails the simplification of the GPR rules. In GECKO 3, reactions catalysed by isozymes (“OR” rules) are separated into different reactions catalysed by a single enzyme. The CORAL Toolbox introduces another simplification, dealing with reactions catalysed by multiple enzymes (“AND” rules), by splitting all reactions catalysed by enzyme complexes into multiple partial reactions catalysed by one enzyme each (Figure 1A).

In the second step, we ordered the reactions catalysed by promiscuous enzymes from the lowest to highest MW/k_{cat} ratio. We defined as the main reaction the

one with the largest k_{cat} value (i.e., lowest MW/k_{cat} ratio); the other reactions catalysed by the same enzyme are considered side reactions. For each reaction catalysed by a promiscuous enzyme, a pseudoenzyme was created to replace the promiscuous enzyme. A number was appended to the IDs of each pseudoenzyme according to the order of MW/k_{cat} ratios. These pseudoenzymes comprises a subpool that draw resources from the enzyme pool, specific to one enzyme. Each enzyme pool then draws resources from the total protein pool, which is obtained from an exchange reaction (Figure 1B).

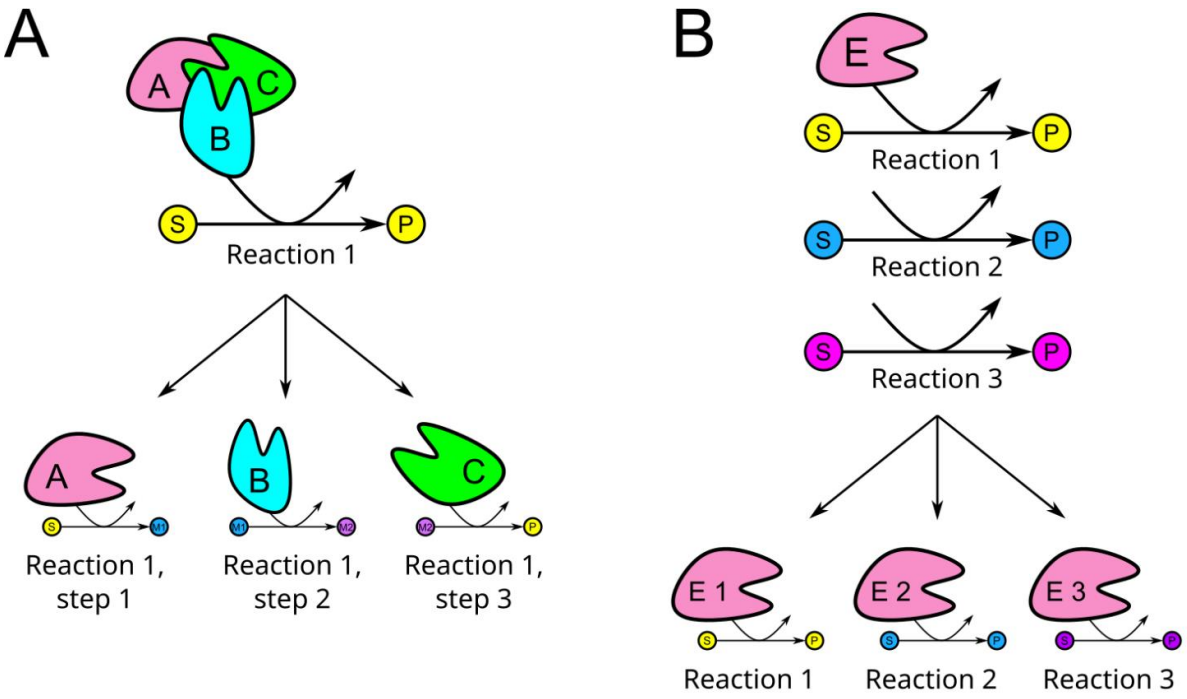

Figure 1. **Restructuring of the pcGEM as performed in CORAL. A)** Simplification of GPR rules to split reactions catalysed by enzyme complexes into multiple partial reactions. **B)** Splitting of promiscuous enzymes into subpools of enzymes, each catalysing a single reaction

Lastly, GECKO 3 introduces a new structure in the YAML and MAT file formats, the ".ec" structure, where all enzyme information is stored. Given that CORAL changes how enzymes are integrated in the model, most of this information no longer matches the indexes of reactions, metabolites, proteins and genes that is present in the model

after the changes take place. The third step updates all enzyme information and makes it available at the new ".und" structure, while the original ".ec" structure remains available.

Flux variability analysis

To evaluate the ranges of metabolic fluxes and of enzyme usages in the CORAL model, we performed flux variability analysis (FVA). We first maximized and minimized the flux v of reaction j to find its flux range:

$$\max_{\mathbf{v}} / \min_{\mathbf{v}} v_j \tag{P1}$$

subject to:

$$\mathbf{S} \cdot \mathbf{v} = 0 \tag{1}$$

$$\mathbf{v_{min}} \leq \mathbf{v} \leq \mathbf{v_{max}} \tag{2}$$

$$v_j \leq k_{cat}^{i,j} \cdot E_{s,i} \tag{3}$$

$$\frac{E_{s,i}}{E_{s,m}} = \frac{k_{cat}^{m,j}}{k_{cat}^{i,j}} \tag{4}$$

$$\sum E_{s,i} = E_s \tag{5}$$

$$\sum E_s = E_{tot} \tag{6}$$

where $\mathbf{S}$ is the stoichiometric matrix, $\mathbf{v}$ is the flux distribution vector, $\mathbf{v_{min}}$ is the flux lower bound vector, $\mathbf{v_{max}}$ is the flux upper bound vector, $k_{cat}^{i,j}$ is the k_{cat} value for the enzyme subpool i, catalysing the reaction j, E is the enzyme usage in the enzyme s, subpool i, $E_{s,m}$ is the enzyme usage for any subpool m that is not catalysing the main reaction, $k_{cat}^{m,n}$ is the k_{cat} value for enzyme m, and E_{tot} is the total protein pool. The ratio constraint introduces a proportional relationship where more inefficient enzymes (i.e., lower k_{cat} values) would require a higher allocation of enzyme resources to carry

the same amount of flux. In other words, we aim to model the prioritization of enzyme resources based on catalytic efficiency of the enzyme. Next, we performed FVA on the enzyme usage pseudoreactions using the same constraints as before, but with the following objective function:

$$\max_{E}/\min_{E} E_{s,i} \qquad \text{(P2)}$$

Single and double knockouts of subpools

We investigated how enzyme resources are distributed to side reactions when the main reaction is knocked out. To this end, we first used a modified implementation of pFBA where we minimize the sum of enzyme subpools instead of sum of fluxes. In the first round we obtain the enzyme subpool distribution we later use in a second round:

$$\min_{E} \mathbf{E}^{\mathbf{subpools}} = \sum_{s=1}^{n} E_s \qquad \text{(P3)}$$

subject to:

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

$$\mathbf{v_{min}} \leq \mathbf{v} \leq \mathbf{v_{max}} \qquad (2)$$

$$v_j \leq k_{cat}^{i,j} \cdot E_{s,i} \qquad (3)$$

$$\frac{E_{s,i}}{E_{s,m}} = \frac{k_{cat}^{m,j}}{k_{cat}^{i,j}} \qquad (4)$$

$$\sum E_{s,i} = E_s \qquad (5)$$

$$\sum E_s = E_{tot} \qquad (6)$$

$$v_{bio} = \mu \qquad (7)$$

where n is the number of enzymes, v_{bio} is the flux through the biomass pseudoreaction, and μ is the specific growth rate.

After finding the enzyme subpool distribution, it is next used to constrain the model (allowing 5% flexibility) in a second round of simulation, where we set to zero the subpool $E_{s,1}$ of each enzyme s one by one, and optimize growth:

$$\max_{v} \; v_{bio} \qquad \text{(P4)}$$

subject to:

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

$$\mathbf{v_{min}} \leq \mathbf{v} \leq \mathbf{v_{max}} \qquad (2)$$

$$v_j \leq k_{cat}^{i,j} \cdot E_{s,i} \qquad (3)$$

$$\frac{E_{s,i}}{E_{s,m}} = \frac{k_{cat}^{m,j}}{k_{cat}^{i,j}} \qquad (4)$$

$$\sum E_{s,i} = E_s \qquad (5)$$

$$\sum E_s = E_{tot} \qquad (6)$$

$$E_s^{subpools} \cdot 0.95 \leq E_s \leq E_s^{subpools} \cdot 1 \qquad (8)$$

$$E_{s,1} = 0 \qquad (9)$$

where $E_s^{subpools}$ is the enzyme subpool usage for subpool s as predicted in the previous step.

Next, we investigated how knocking out the main reaction impacts growth. To this end, we simulated a new round of knockouts starting from the single knockout P4 problem but excluding the constraints on subpools from the previous solutions:

$$\max_{\mathbf{v}} \; v_{bio} \qquad \text{(P5)}$$

subject to:

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

$$\mathbf{v_{min}} \leq \mathbf{v} \leq \mathbf{v_{max}} \qquad (2)$$

$$v_j \leq k_{cat}^{i,j} \cdot E_{s,i} \quad (3)$$

$$\frac{E_{s,i}}{E_{s,m}} = \frac{k_{cat}^{m,j}}{k_{cat}^{i,j}} \quad (4)$$

$$\sum E_{s,i} = E_s \quad (5)$$

$$\sum E_s = E_{tot} \quad (6)$$

$$E_{s,j} = 0 \quad (10)$$

where the subpool $E_{s,j}$ is the subpool j from any enzyme s. Then, for all single knock-outs that impacted growth, we performed a pairwise double knockout of these subpools as one element of the pair and any other subpool as the other element of the pair to assess how they impact growth:

$$\max_{\mathbf{v}} \ v_{bio} \qquad \text{(P6)}$$

subject to:

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

$$\mathbf{v_{min}} \leq \mathbf{v} \leq \mathbf{v_{max}} \quad (2)$$

$$v_j \leq k_{cat}^{i,j} \cdot E_{s,i} \quad (3)$$

$$\frac{E_{s,i}}{E_{s,m}} = \frac{k_{cat}^{m,j}}{k_{cat}^{i,j}} \quad (4)$$

$$\sum E_{s,i} = E_s \quad (5)$$

$$\sum E_s = E_{tot} \quad (6)$$

$$E_{s,j} = 0 \quad (10)$$

$$E_{d,j} = 0 \quad (11)$$

where the subpool $E_{d,j}$ is the subpool j from any enzyme d that is not the same as the enzyme s. To further assess the effect of knockouts on growth, we performed double knockouts in the conventional GEM, iML1515u, deleting pairs of genes following classical GPR Boolean implementations and optimizing growth as in P6, minus the enzyme constraints.

Lastly, we were also interested in how knocking out other subpools, besides $E_{s,1}$, affect growth. We performed a double knockout of subpools from $E_{s,2}$ to $E_{s,5}$, as past $E_{s,5}$ there are much fewer enzymes with this many subpools.

Flux-sum analysis

To further assess the capabilities of the CORAL model, we investigated the metabolite turnover (flux-sum analysis (CHUNG; LEE, 2009)) across the optimization problems we solved previously. This ensured that not only we take an enzyme- and reaction-centric overview, but also a metabolite-centric assessment. To calculate flux-sums, we apply the following equation:

$$\phi_i = 0.5 \sum_j \left| S_{i,j} v_j \right| \tag{12}$$

where ϕ_i is the flux-sum of metabolite i, $S_{i,j}$ is the stoichiometric coefficient of metabolite i participating in reaction j, and v_j is the flux through reaction j.

Results and discussion

CORAL accounts for promiscuous enzyme activity and underground metabolism

By building upon existing protein-constrained approaches, we developed CORAL as a toolbox to investigate promiscuous enzyme activity and underground metabolism in the context of constraint-based modelling. To this end, we first included underground

reactions into the *E. coli* iML1515 model, resulting in the iML1515u model. Then, using the DLKcat-predicted k_{cat} values, we used GECKO 3 to integrate enzyme constraints into iML1515u. The pcGEM was then restructured using CORAL, and the resulting model is named eciML1515u. In CORAL, the enzyme usage in reactions is restructured in a manner that allows for modelling resource usage in promiscuous enzymes. This is achieved by splitting the enzyme pool for each promiscuous enzyme into as many subpools as there are side reactions (Figure 1). The sum of subpools for a certain enzyme then corresponds to the original enzyme pool (see Equation 5). Before the restructuring, the model contained 3774 metabolites, 8331 reactions, and 1526 enzymes. After the restructuring, eciML1515u now has 12048 metabolites, 16605 reactions, 1526 enzymes, and 7260 subpools (Table 1). The large number of metabolites and reactions is due to the number of pseudometabolites and pseudoreactions added to accommodate for the subpools and their usage of the enzyme pools, along with simplification of GPR rules to split enzyme complexes into partial reactions.

Table 1. General descriptors for the models used in this study.

	iML1515	**iML1515u**	**eciML1515u (pre-CORAL)**	**eciML1515u (CORAL)**
Genes	1516	1530	1531	1531
Reactions	2712	3238	8331	16605
Metabolites	1877	2247	3774	12048
Enzymes	-	-	1526	1526
Enzyme subpools	-	-	-	7260

Promiscuous enzymes increase metabolic flux variability

Promiscuous enzyme activity and underground metabolism provides alternative flux routes. To investigate the impact of these added reactions, we performed FVA using

eciML1515u, both with and without underground reactions. We further tested how chemostat conditions affect flux variability by using a fixed growth rate. We found that flux variability is higher when underground reactions are present (Figure 2A). The scenario without underground reactions and no fixed growth rate had a lower flux variability in 79.85% of reactions when compared to the condition with underground reactions and no fixed growth rate. The scenario with fixed growth rate and without considering underground reactions showed lower flux variability in 79.22% of reactions in comparison to the condition with underground reactions and fixed growth rate. Therefore, we concluded, in line with the expectation, that underground metabolism leads to higher variability of metabolic fluxes.

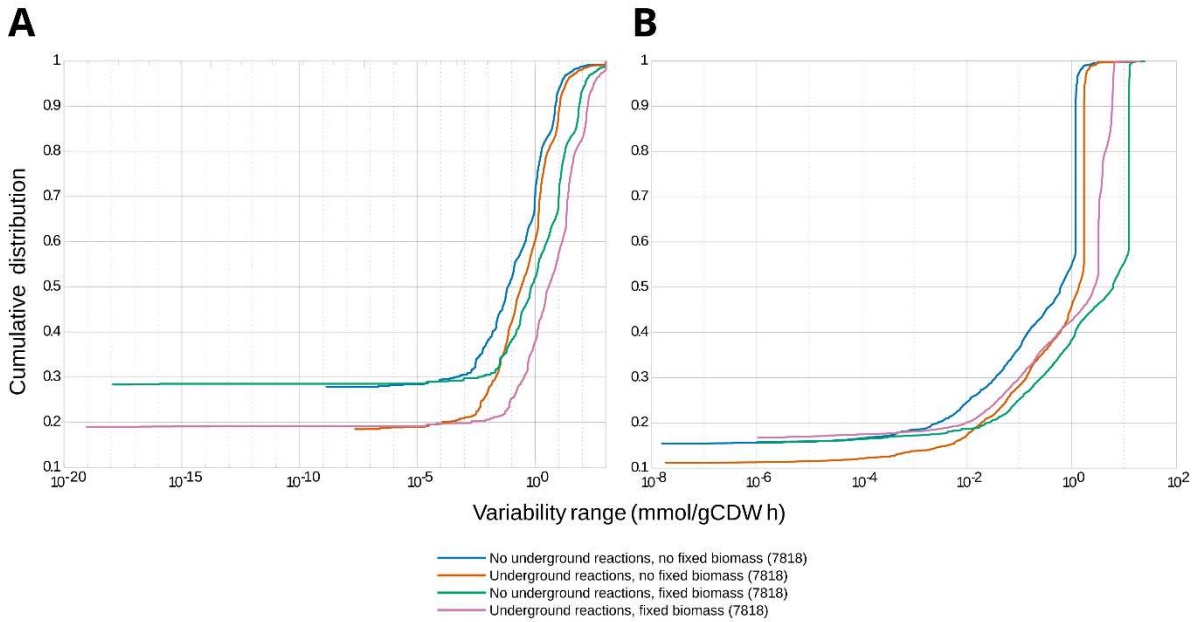

Figure 2. **Flux variability for the CORAL-restructured model**. We investigated flux and resource allocation variability with or without underground reactions and with or without fixed biomass, as indicated in the legend. **A)** Flux variability for metabolic reactions only (excluding all GECKO- and CORAL-related pseudoreactions). **B)** Flux variability for enzyme subpool usage.

Next, we performed FVA to check the variability of subpool usage instead of metabolic flux. We found that subpool usage variability behaves similarly to metabolic

flux variability, with underground reactions leading to an increase in the overall variability (Figure 2B). The subpool usage range was larger in 82.13% of subpools for the condition with underground reactions and no fixed growth rate. Similarly, the subpool usage was larger in 83.30% of subpools for the condition with underground reactions and a fixed growth rate.

Redistribution of enzyme resources to side reactions following knockout of a main reaction ensures metabolic robustness

The restructuring of enzyme usage in CORAL allows for a closer inspection of resource allocation, as promiscuous enzymes are now represented as different pseudometabolites for each reaction it catalyses. Given that, biologically, the larger part of enzyme resources is allocated to the main reaction (SINGLA; BHARDWAJ, 2020), we investigated how these resources are distributed to the side reactions when the main reaction no longer occurs. To achieve this, we simulated a knockout of the main reaction by setting its enzyme subpool to zero, ensuring that there is no enzyme subpool available to catalyse the main reaction. This is necessary as performing knockouts using conventional GPR rules would have impacts on all enzyme subpools of a certain enzyme, along with disturbing additional reactions.

In a first round of simulations, we determine the enzyme subpool usage distribution in the wildtype by solving an optimization problem (P3, Methods). Next, we used the enzyme subpool usage distribution to constrain the enzyme subpool usage distributions in the knock-out mutant that abolishes the main activity. These simulations were performed under the same constraints on substrate uptake rates to favour a respiratory metabolism. We found a total of 38 knockout solutions where the main reaction

enzyme subpool was used in the wildtype solution, and whose knockout resulted in non-lethal mutants.

Considering the changes occurring within an enzyme pool, 13 of the 38 solutions resulted in an almost direct redistribution of $E_{s,1}$ to $E_{s,2}$ (which has the second highest catalytic efficiency), such that the enzyme subpool usage value for $E_{s,2}$ is the same as $E_{s,1}$ in the P3 solution or within 5% proximity (since we allow for 5% flexibility). For 10 other solutions, there is a higher allocation to $E_{s,2}$, while other subpools also receive resources but at a lesser amount. For seven solutions, other enzyme subpools received more resources than $E_{s,2}$. Lastly, in eight solutions, the subpool $E_{s,1}$ is redistributed entirely to enzyme subpools other than $E_{s,2}$, which receives zero.

To assess the magnitude of the changes in enzyme subpool allocation after a knockout of the main reaction subpool, we calculated the ratio of how much each enzyme subpool uses from the total enzyme pool ($E_{s,j}/E_s$). We found that none of the enzyme subpools that showed a significant increase in the mutant compared to the wildtype were used in the wildtype solution. Inspecting the enzyme subpools with the highest ratio after the knockout, the one with the highest increase belong to the enzyme pool of the outer membrane porin C, which has 566 subpools. Its main enzyme subpool catalyses the import of calcium in the model. Before the knockout, the main enzyme subpool takes 98.03% of the enzyme pool, whereas after the knockout, the resources were used instead by the subpool 522, which takes 98.12% of the enzyme pool. The enzyme subpool 522 catalyses the export of L-tryptophan to the environment. Another subpool that significantly takes over the enzyme pool is the subpool 2 of the amino acid acetyltransferase. The subpool 2 catalyses the transfer of the acetyl group from acetyl-CoA into DL-2-aminopimelate, forming 2-acetylaminoheptanedioate and

coenzyme A and taking 83% of the enzyme pool. The list of subpools and their corresponding percentage of the enzyme pool is listed in Supplementary Data 1.

Considering global changes in the model following a knockout, we calculated the difference between enzyme subpool usage predicted for the wildtype and all the mutant solutions. In most solutions, the enzyme subpool usage values have a slight change, with a few enzyme subpools having a larger change (Figure S1). Among the enzyme subpools with the largest change, there are the enzyme subpools 2 and 3 of the subunit alpha of the ATP synthase complex. These enzyme subpools catalyses the inverse reaction of ATP synthase, producing ADP from ATP hydrolysis. Another impacted enzyme pool is the subunit beta of the ATP synthase complex, of which the enzyme subpools 2 and 3 also have a higher increase compared to subpools in other enzyme pools. The enzyme subpools 2 and 3 of the ATP synthase gamma chain are also highly impacted. These enzyme subpools all receive the largest part of the enzyme pool after the knockout of several different main enzyme subpools, suggesting that the ATP/ADP balance is disturbed when main enzyme subpools are knocked out. A lower ATP/ADP ratio leads to oxidative stress and reduction in growth and protein synthesis (SORIA et al., 2024).

Given that the main reaction no longer occurs, we checked if the metabolites involved in these reactions display any change in flux-sums, which can be a useful proxy for metabolite concentrations (CHUNG; LEE, 2009; LAKSHMANAN et al., 2015). We calculated the flux-sums of the wildtype solution and of all 38 mutant solutions and calculated their differences. Surprisingly, we found that very few metabolites display changes in flux-sums (Figure 3). Notably, both NAD and NADH presented reduced flux-sums across P4 solutions. This correlates with the disturbance of the ATP/ADP ratio described earlier, as energy metabolism is dependent on the availability of both

NAD and NADH. The metabolite with increased flux-sum is UDP-*N*-acetyl-D-glucosamine, which is the precursor of peptidoglycan and is involved in the biosynthesis of lipopolysaccharides (GALINIER et al., 2023). Taken together, these results indicate that while a few metabolites are impacted, the cell metabolism is overall robust to disturbances in enzymes with promiscuous activity, suggesting that having alternative routes can mitigate the impacts of knocking out main reactions.

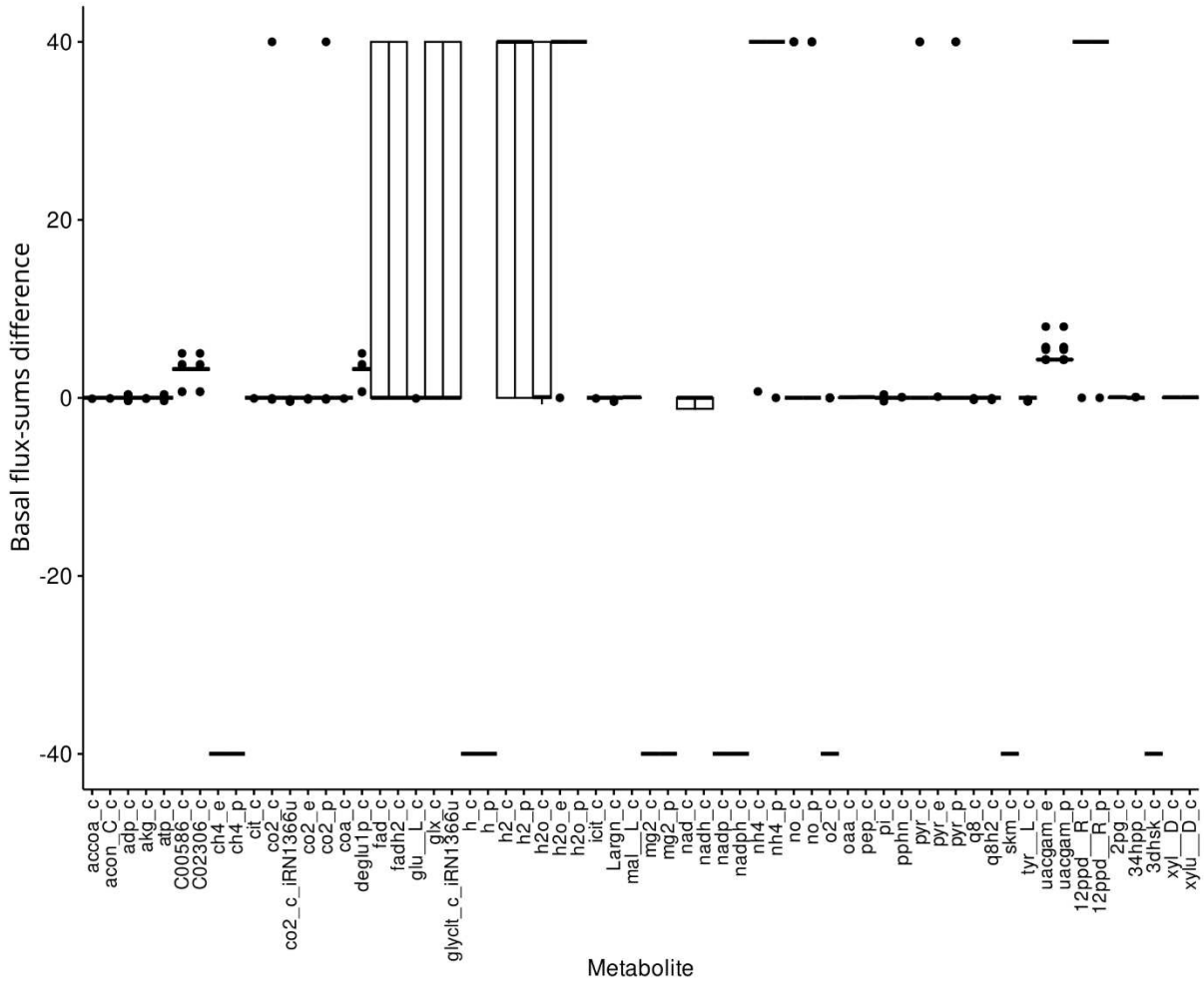

Figure 3. **Difference in flux-sums between the wildtype and the mutant solutions.** We calculated the flux sums only for metabolites that were common to all solutions when knocking out a single main reaction. We calculated the values by subtracting the flux-sums of the wildtype solution from the flux-sums of the mutant solutions

Double knockouts of pairs of enzymes with no promiscuous activity have higher impact on growth

We next assessed how growth is impacted when pairs of main reactions are blocked. To this end, we knocked out a pairwise combination of all enzyme subpools that catalysed main reactions (e.g., $E_{1,1}$ and $E_{2,1}$) and compared how it changes compared to a growth rate of 0.11 h^{-1} for the wild type (WT). Comparing to this result, we obtained 3309 pairs out of 24222 whose knockout reduced this growth rate by least 1%. The enzyme subpool pairs used a combination of 156 different enzymes, of which 68 enzymes have promiscuous activity in the model.

Taking a closer look at the knockout pairs that most impacted growth (higher than 1% of the growth rate), two enzymes stand out as causing the growth rate to drop no matter the other component of the pair, with the 618 most impacted pairs out of 3309 containing either of these two enzymes. These are the inosine-5'-monophosphate dehydrogenase, whose main reaction in the model catalyses the conversion of inosine-5'-phosphate (IMP) to xanthosine-5'-phosphate (XMP); and the probable acyl-CoA dehydrogenase YdiO, whose main reaction in the model catalyses the conversion of crotonoyl-CoA into butanoyl-CoA using one reduced FAD. Further, these two enzymes form the pair that most reduces the growth rate, with the double knockout to wild-type growth rate ratio of 0.57. The metabolite xanthosine-5'-phosphate is an intermediary of the *de novo* synthesis of guanine nucleotides, thus being important for growth. Likewise, butanoyl-CoA is an intermediate metabolite in fatty acid metabolism, in both β-oxidation and elongation in mitochondria. In the CORAL model, these enzymes have no promiscuous activity, meaning they have no enzyme subpools to catalyse reactions different than the main reaction.

Amongst the 20912 enzyme pairs that caused no impact on growth (equal or less than 1% of the growth rate), in 15052 pairs there is at least one enzyme with promiscuous activity (Figure 4A), meaning these have resources redistributed to other enzyme subpools upon knocking out the main reaction enzyme subpool. Meanwhile, in the pairs that affect growth, there is a higher prevalence of enzymes without promiscuous activity (Figure 4B). The table containing all knockout pairs is available in Supplementary Data 2.

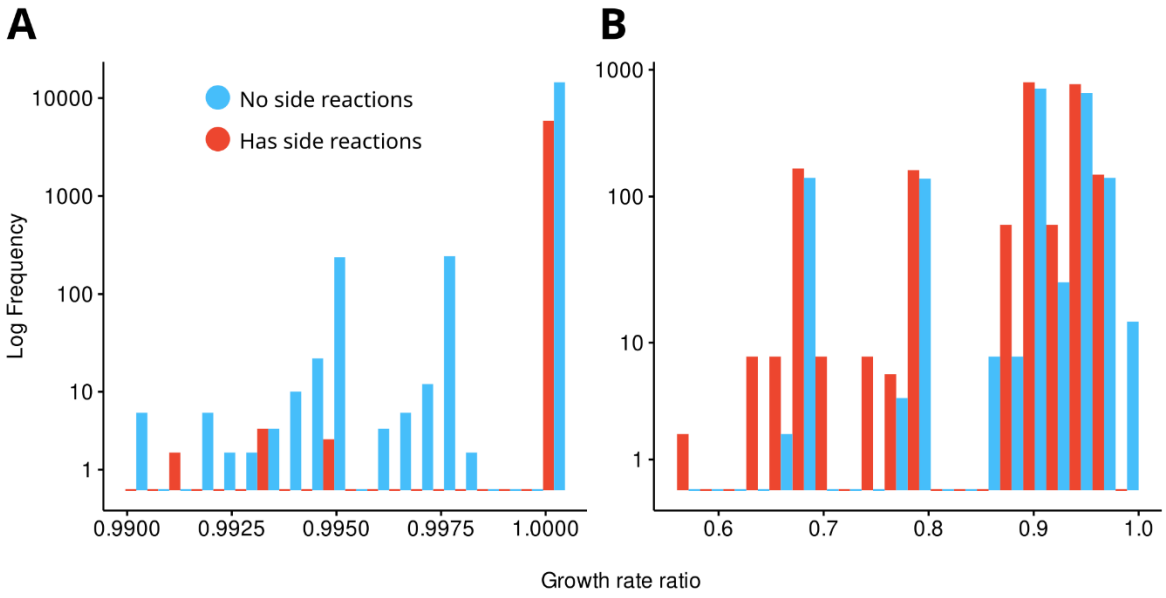

Figure 4. Distribution of growth rate ratios ($^{\mu_{WT}}/_{\mu_{del}}$) for double knockout mutant solutions.

A) Growth rate ratios obtained from main reaction pairs that did not affect growth or affected it by less than 1%. **B)** Growth rate ratios obtained from main reaction pairs that affected growth by at least 1%.

To further assess the effects of the double knockouts on growth, we also deleted pairs of genes from the conventional GEM iML1515u (Figure 5A). We knocked out pairs of genes following GPR rules for all reactions. This is useful to assess how much enzyme promiscuity matters for maintaining the growth rate, since now all reactions controlled by the knocked-out gene will no longer occur. In this scenario, we find that

this indeed results in lower growth rates for pairs of genes that code for promiscuous enzymes when no side reactions are available (Figure 5B), suggesting that when all reactions, main or side, catalysed by an enzyme are blocked at the same time, the compensating effect observed in the CORAL model is lost.

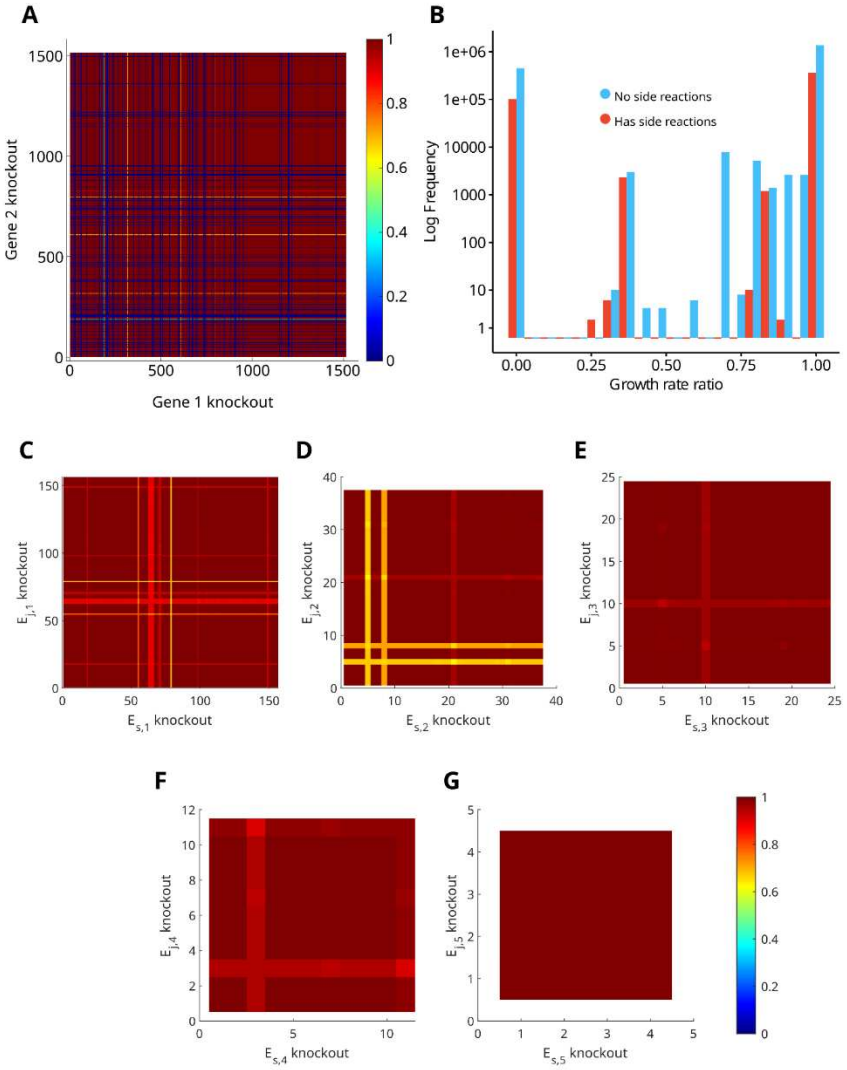

Figure 5. **Impact on growth after double knockouts. A)** Impact on growth following double knockouts in the iML1515u model, performed by deleting pairs of genes according to GPR rules. **B)** Distribution of growth rate ratios following double knockouts in the iML1515u model. **C)** Double knockout of first subpool ($E_{s,1}$) in the CORAL model. **D)** Second subpool ($E_{s,2}$). **E)** Third subpool ($E_{s,3}$). **F)** Fourth subpool ($E_{s,4}$). **G)** Fifth subpool ($E_{s,5}$).

Altogether, these results corroborate the findings that promiscuous reactions are important for the maintenance of the growth rate as well as mitigating impacts on cell metabolism. We also inspected what happens when we knock out any of the enzyme subpools from $E_{s,2}$ to $E_{s,5}$, and noticed that there is little impact on the growth rate. When knocking out the enzyme subpool $E_{s,2}$ (Figure 5C), there were four enzymes whose presence in a pair had the most impact had the growth rate ratio (≤ 0.71). For the enzyme subpools $E_{s,3}$ to $E_{s,5}$, the lowest growth rate ratio was 0.905 (Figure 5C-G). In contrast, knockout pairs of main enzyme subpool had significantly more impact on the growth rate. This suggests that while side reactions are essential to maintain metabolic and growth balance when the main reaction is impacted, in the opposite situation (knocking out side reactions while enabling the main reaction), the double knockouts have negligible impact on growth. The same is observed when different carbon sources are evaluated (Figure S2).

Conclusion

Here we present the CORAL Toolbox, a tool for the integration and analysis of promiscuous enzyme activity and underground metabolism. We demonstrated that the inclusion of underground reactions increases metabolic flux variability. Furthermore, the redistribution of enzyme resources from the main to the side reactions highlights the importance of these reactions to maintain metabolic flexibility and robustness, serving as alternative routes that compensates for the loss of the main reaction. Moreover, when inspecting the impact of double knockouts of main reactions on growth, promiscuous enzyme activity and underground metabolism proved essential to maintaining growth, as the knockouts with the highest impact were on enzymes without promiscuity. Further, when simulating a scenario where promiscuous enzyme activity does not

take place, we see that this compensating effect is lost. A relaxation of the equality constraint (Eq. 4) into an inequality could be assessed to check if the conclusions hold in a relaxed model. These results highlight the importance of considering underground metabolism and enzyme promiscuity in terms of metabolic plasticity in constraint-based approaches, allowing for a comprehensive understanding of metabolism and its adaptation to disturbances, pinpointing that flexibility is crucial for cell survival and resilience.

Acknowledgements

We thank Marius Arend, Philipp Wendering and Fayaz Soleymani for their critical discussion and comments on this study.

Data availability

The CORAL Toolbox is publicly available in a GitHub repository along with the code and data for reproducing this work: https://github.com/mauricioamf/CORAL.

References

ADADI, R. et al. Prediction of Microbial Growth Rate versus Biomass Yield by a Metabolic Network with Kinetic Parameters. **PLOS Computational Biology**, v. 8, n. 7, p. e1002575, jul. 2012.

CHEN, Y. et al. Reconstruction, simulation and analysis of enzyme-constrained metabolic models using GECKO Toolbox 3.0. **Nature Protocols**, v. 19, n. 3, p. 629–667, mar. 2024.

CHUNG, B. K. S.; LEE, D.-Y. Flux-sum analysis: a metabolite-centric approach for understanding the metabolic network. **BMC Systems Biology**, v. 3, n. 1, p. 1–10, dez. 2009.

COPLEY, S. D. An evolutionary biochemist's perspective on promiscuity. **Trends in Biochemical Sciences**, v. 40, n. 2, p. 72–78, 1 fev. 2015.

COPLEY, S. D. Shining a light on enzyme promiscuity. **Current Opinion in Structural Biology**, Protein–nucleic acid interactions • Catalysis and regulation. v. 47, p. 167–175, 1 dez. 2017.

COPLEY, S. D.; NEWTON, M. S.; WIDNEY, K. A. How to Recruit a Promiscuous Enzyme to Serve a New Function. **Biochemistry**, v. 62, n. 2, p. 300–308, 17 jan. 2023.

D'ARI, R.; CASADESÚS, J. Underground metabolism. **BioEssays**, v. 20, n. 2, p. 181–186, 1998.

DINIZ, R. H. S. et al. Transcriptome analysis of the thermotolerant yeast *Kluyveromyces marxianus* CCT 7735 under ethanol stress. **Applied Microbiology and Biotechnology**, v. 101, n. 18, p. 6969–6980, 2017.

GALINIER, A. et al. Recent Advances in Peptidoglycan Synthesis and Regulation in Bacteria. **Biomolecules**, v. 13, n. 5, p. 720, 22 abr. 2023.

GLASNER, M. E.; TRUONG, D. P.; MORSE, B. C. How enzyme promiscuity and horizontal gene transfer contribute to metabolic innovation. **The FEBS Journal**, v. 287, n. 7, p. 1323–1342, 2020.

GUZMÁN, G. I. et al. Enzyme promiscuity shapes adaptation to novel growth substrates. **Molecular Systems Biology**, v. 15, n. 4, p. e8462, abr. 2019.

HEIRENDT, L. et al. Creation and analysis of biochemical constraint-based models using the COBRA Toolbox v.3.0. **Nature Protocols**, v. 14, n. 3, p. 639–702, 2019.

KOVÁCS, S. C. et al. Underground metabolism as a rich reservoir for pathway engineering. **Bioinformatics**, v. 38, n. 11, p. 3070–3077, 26 maio 2022.

LAKSHMANAN, M. et al. Flux-sum analysis identifies metabolite targets for strain improvement. **BMC Systems Biology**, v. 9, n. 1, p. 1–11, dez. 2015.

LI, F. et al. Deep learning-based kcat prediction enables improved enzyme-constrained model reconstruction. **Nature Catalysis**, p. 1–11, 16 jun. 2022.

MO, W. et al. *Kluyveromyces marxianus* developing ethanol tolerance during adaptive evolution with significant improvements of multiple pathways. **Biotechnology for Biofuels**, v. 12, n. 1, p. 1–15, 2019.

MONK, J. M. et al. iML1515, a knowledgebase that computes *Escherichia coli* traits. **Nature Biotechnology**, v. 35, n. 10, p. 8–12, 2017.

NOTEBAART, R. A. et al. Network-level architecture and the evolutionary potential of underground metabolism. **Proceedings of the National Academy of Sciences**, v. 111, n. 32, p. 11762–11767, 12 ago. 2014.

OBERHARDT, M. A. et al. Systems-Wide Prediction of Enzyme Promiscuity Reveals a New Underground Alternative Route for Pyridoxal 5'-Phosphate Production in E. coli. **PLOS Computational Biology**, v. 12, n. 1, p. e1004705, 28 jan. 2016.

ORTH, J. D. et al. A comprehensive genome-scale reconstruction of *Escherichia coli* metabolism—2011. **Molecular Systems Biology**, v. 7, p. 535, 11 out. 2011.

ORTH, J. D.; THIELE, I.; PALSSON, B. Ø. What is flux balance analysis? **Nature Biotechnology**, v. 28, n. 3, p. 245–248, 1 mar. 2010.

PAN, S.; REED, J. L. Advances in gap-filling genome-scale metabolic models and model-driven experiments lead to novel metabolic discoveries. **Current Opinion in Biotechnology**, v. 51, p. 103–108, jun. 2018.

SÁNCHEZ, B. J. et al. Improving the phenotype predictions of a yeast genome-scale metabolic model by incorporating enzymatic constraints. **Molecular Systems Biology**, v. 13, n. 8, p. 935, 1 ago. 2017.

SCHMIDT, A. et al. The quantitative and condition-dependent *Escherichia coli* proteome. **Nature Biotechnology**, v. 34, n. 1, p. 104–110, 2016.

SINGLA, P.; BHARDWAJ, R. D. Enzyme promiscuity – A light on the “darker” side of enzyme specificity. **Biocatalysis and Biotransformation**, v. 38, n. 2, p. 81–92, 3 mar. 2020.

SORIA, S. et al. Transcriptional and Metabolic Response of a Strain of *Escherichia coli* PTS− to a Perturbation of the Energetic Level by Modification of [ATP]/[ADP] Ratio. **BioTech**, v. 13, n. 2, p. 10, 10 abr. 2024.

TAWFIK, O. K. AND D. S. Enzyme Promiscuity: A Mechanistic and Evolutionary Perspective. **Annual Review of Biochemistry**, v. 79, n. Volume 79, 2010, p. 471–505, 7 jul. 2010.

VALGEPEA, K. et al. Escherichia coli achieves faster growth by increasing catalytic and translation rates of proteins. **Molecular BioSystems**, v. 9, n. 9, p. 2344–2358, 30 jul. 2013.

WANG, H. et al. RAVEN 2.0: A versatile toolbox for metabolic network reconstruction and a case study on Streptomyces coelicolor. **PLOS Computational Biology**, v. 14, n. 10, p. e1006541, 18 out. 2018.

WENDERING, P.; NIKOLOSKI, Z. Genome-Scale Modeling Specifies the Metabolic Capabilities of Rhizophagus irregularis. **mSystems**, v. 7, n. 1, 22 fev. 2022.

APPENDIX A – SUPPLEMENTARY MATERIAL FOR CHAPTER 2.2

Supplementary material as included in the preprint version.

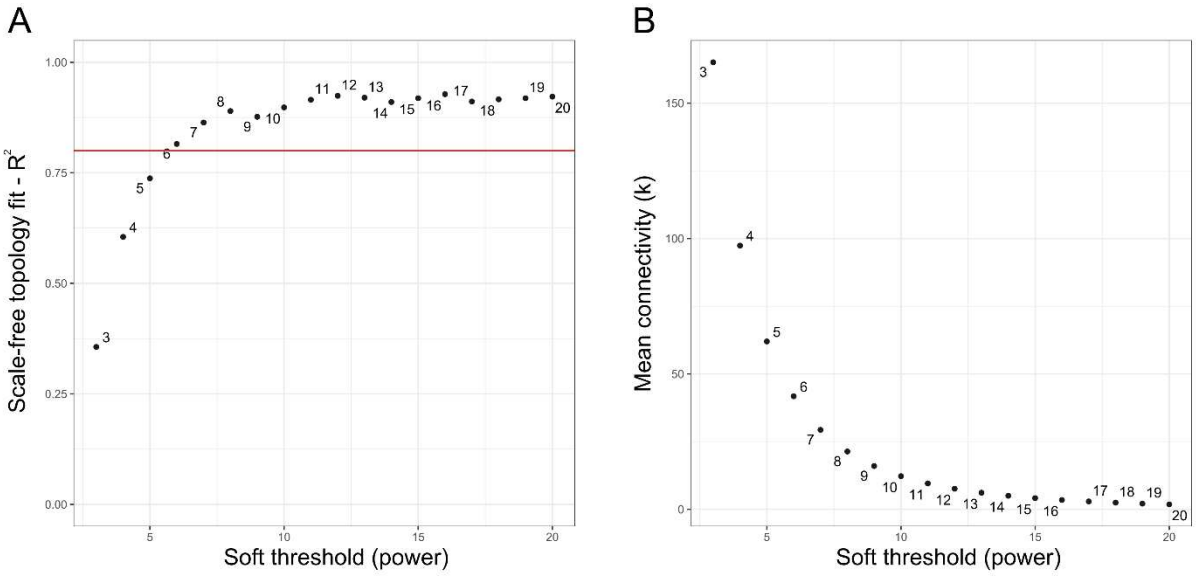

Figure S1. Relationship between scale-free topology and average connectivity for the data from Diniz et al. (DINIZ et al., 2017). The threshold value of 6 was chosen for the subsequent analyses (A) Scale independence. (B) Average connectivity.

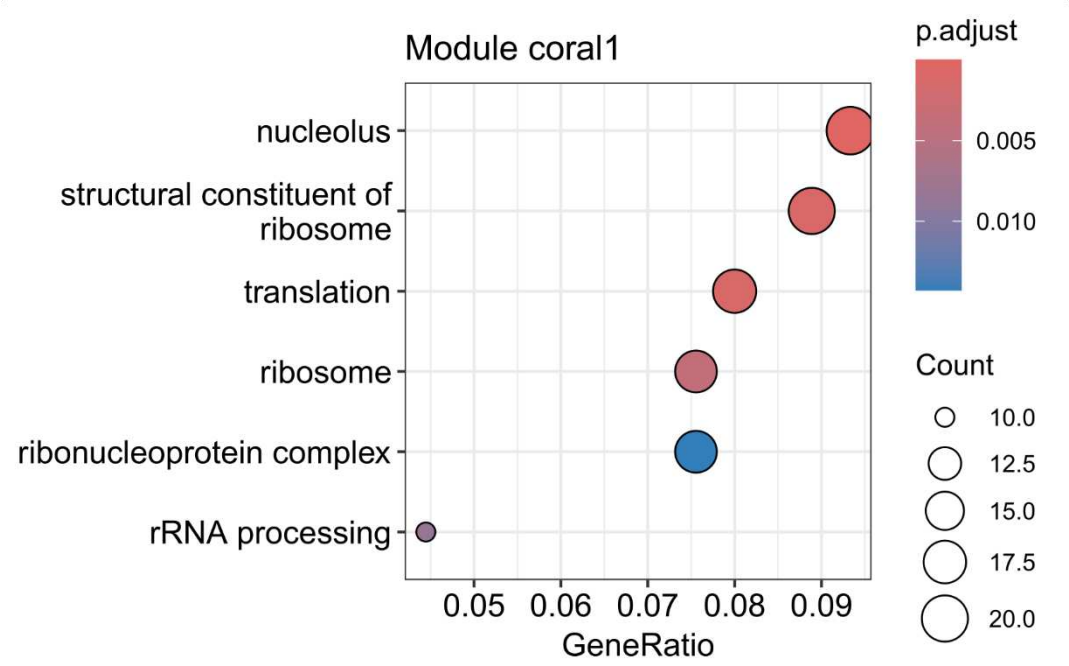

Figure S2. GO enrichment for the coral1 module, the most correlated module in the 0h condition of Diniz et al. (DINIZ et al., 2017, p. 201).

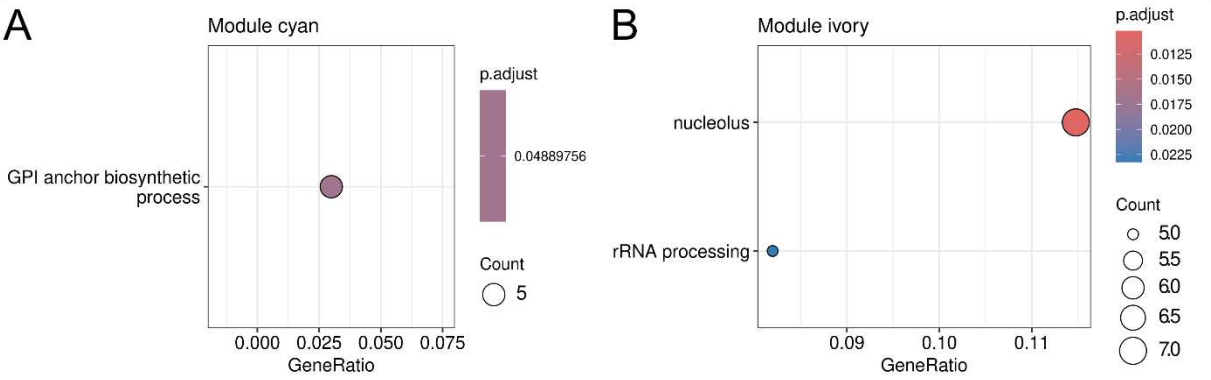

Figure S3. GO enrichment for the cyan (A) and ivory (B) modules, modules most correlated to the 1h condition of Diniz et al. (DINIZ et al., 2017).

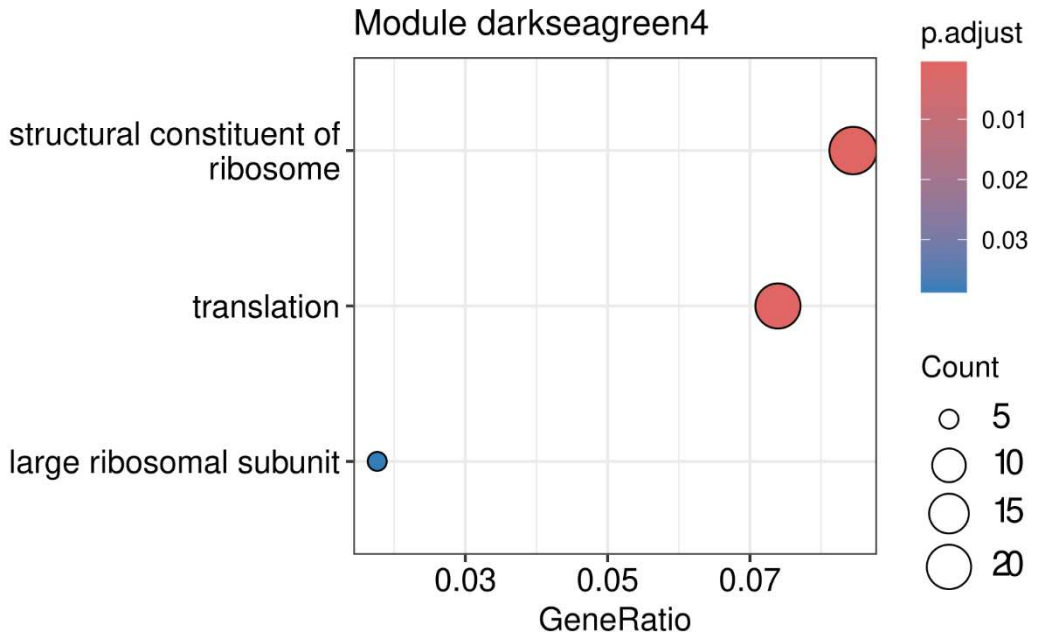

Figure S4. GO enrichment for the darkseagreen4 module, the module most correlated to the 4h condition of Diniz et al. (DINIZ et al., 2017).

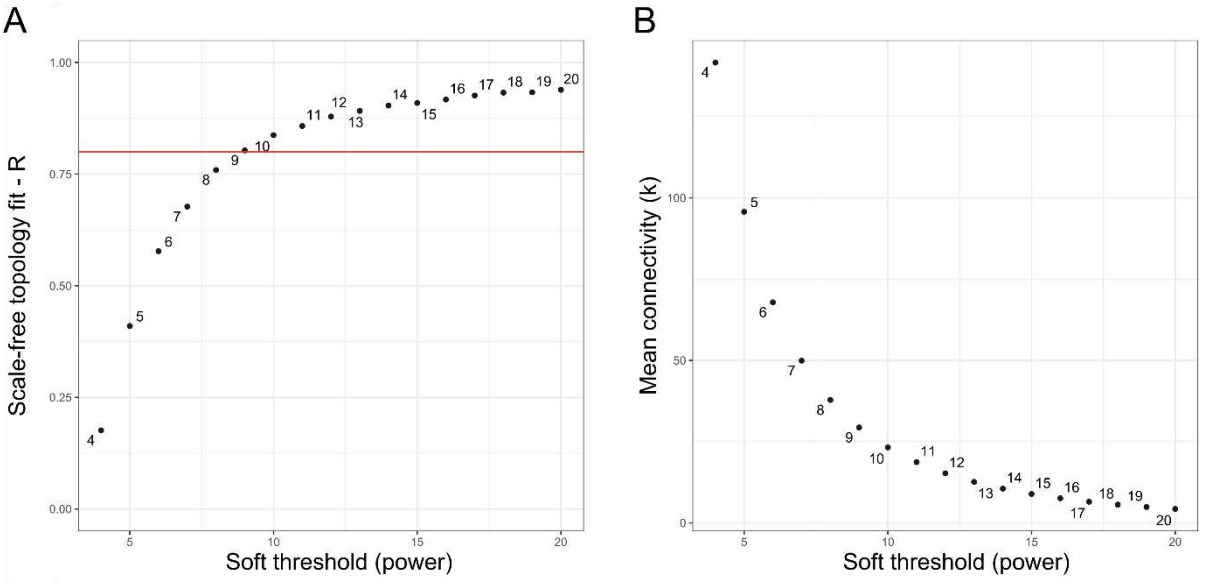

Figure S5. Relationship between scale-free topology and average connectivity for the data from Mo et al. (MO et al., 2019). The threshold value of 9 was chosen for the subsequent analyses (A) Scale independence. (B) Average connectivity.

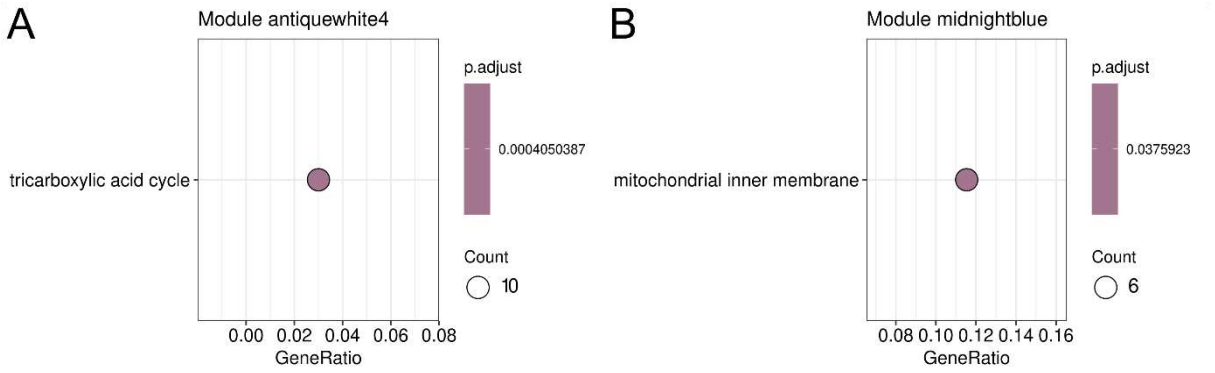

Figure S6. GO enrichment for the antiquewhite4 and midnightblue modules, the most correlated modules in the 4%-KM condition of Mo et al. (MO et al., 2019).

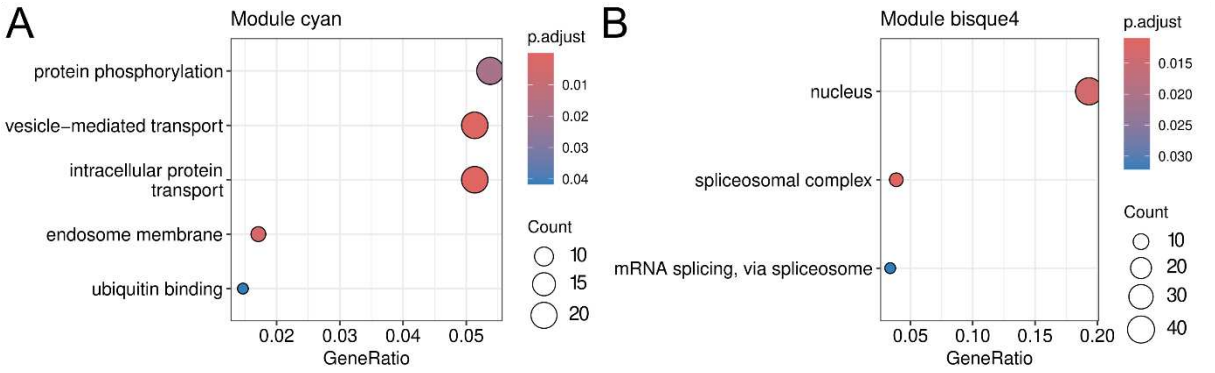

Figure S7. GO enrichment for the cyan (A) and bisque4 (B) modules, the most correlated modules in the 6%-KM (A and B) and 0%-100d (A) condition from Mo et al. (MO et al., 2019).

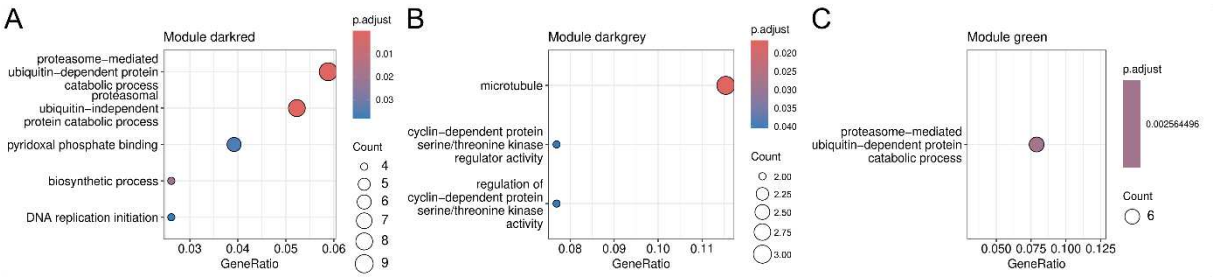

Figure S8. GO enrichment for the darkred (A), darkgrey (B) and green (C) modules, the most correlated modules in the 4%-100d condition of Mo et al. (MO et al., 2019).

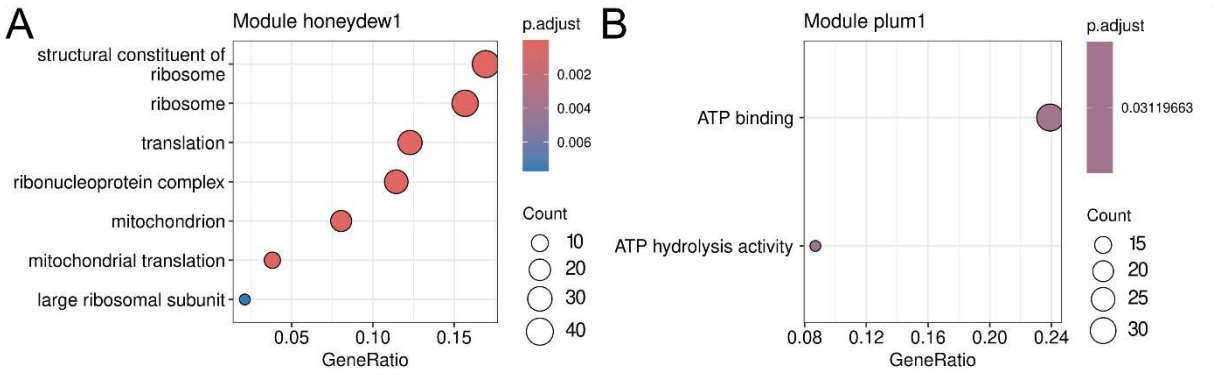

Figure S9. GO enrichment for honeydew1 (A) and plum1 (B), the most correlated modules in the 6%-100d condition from Mo et al. (MO et al., 2019).

Table S1. Growth rates obtained from knocking out regulators in the Diniz et al. [39] data.

Regulator	**Growth rate (1/h)**
0h	
KLMA_20797	3,65E-08
KLMA_40311	2,17E-29
KLMA_70361	2,17E-29
KLMA_80115	2,74E-05
1h	
KLMA_10029	0,147658249
KLMA_40232	0,098438833
KLMA_60332	2,17E-29
KLMA_80190	1,61E-06
4h	
KLMA_10330	1,61E-06
KLMA_20036	1,61E-06
KLMA_20565	0
KLMA_70081	1,61E-06
KLMA_70361	4,08E-23
KLMA_80236	2,17E-29

Table S2. Growth rates obtained from knocking out regulators in the Mo et al. [40] data. (available as an Excel spreadsheet at the GitHub repository: https://github.com/LabFisUFV/KmarxianusEthanol).

APPENDIX B – SUPPLEMENTARY MATERIAL FOR CHAPTER 3.2

Supplementary material as available in the publication.

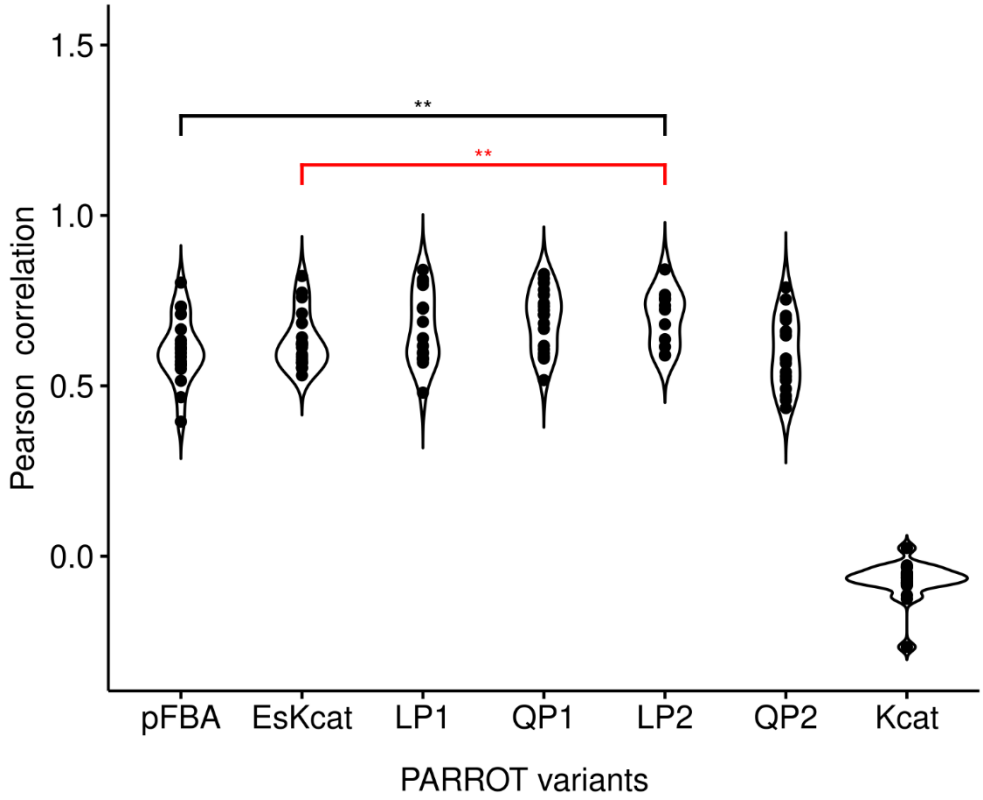

Fig. S1. Pearson correlation calculated between predicted enzyme distribution and the baseline obtained from minimizing the 2-norm of the experimental enzyme usage distribution, in *S. cerevisiae*. All values were log10-transformed prior to comparisons. A pairwise Wilcoxon rank sum assesses the statistical significance: ** p-value < 0.0009. Black significance bar indicates comparisons to pFBA. Red significance bar indicates comparisons to EsKcat.

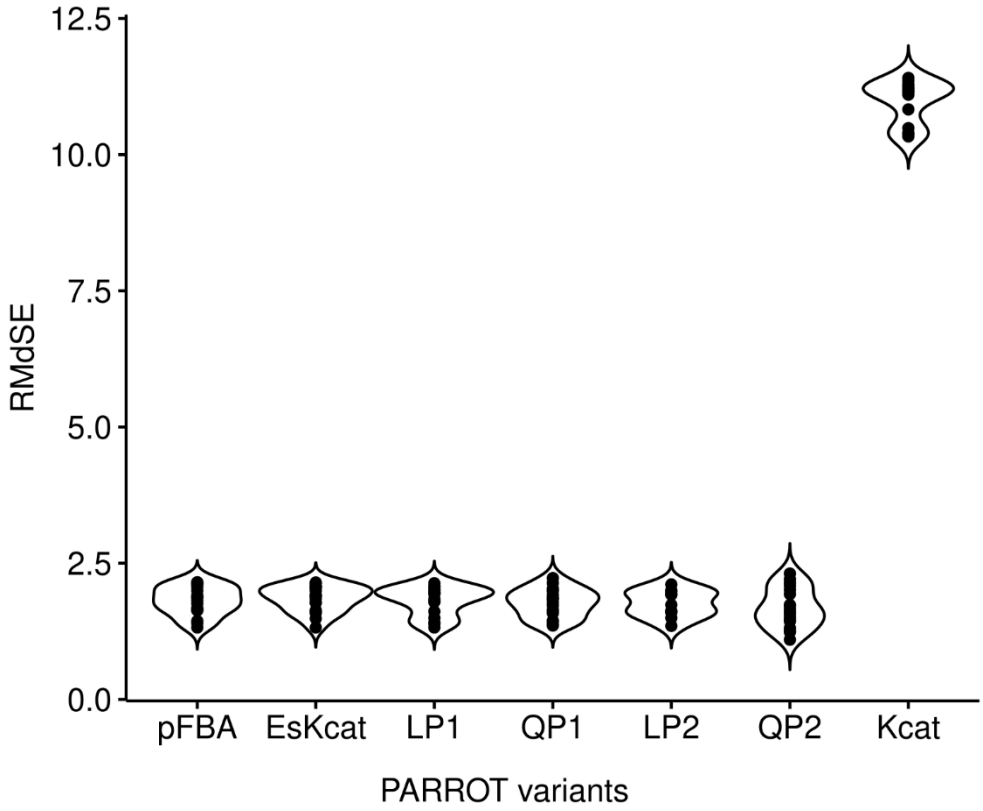

Fig. S2. Assessment of model performance based on the root median squared error (RMdSE). The minimization of the 2-norm of the experimental enzyme usage distribution in *S. cerevisiae* was used. All values were log10-transformed prior to comparisons.

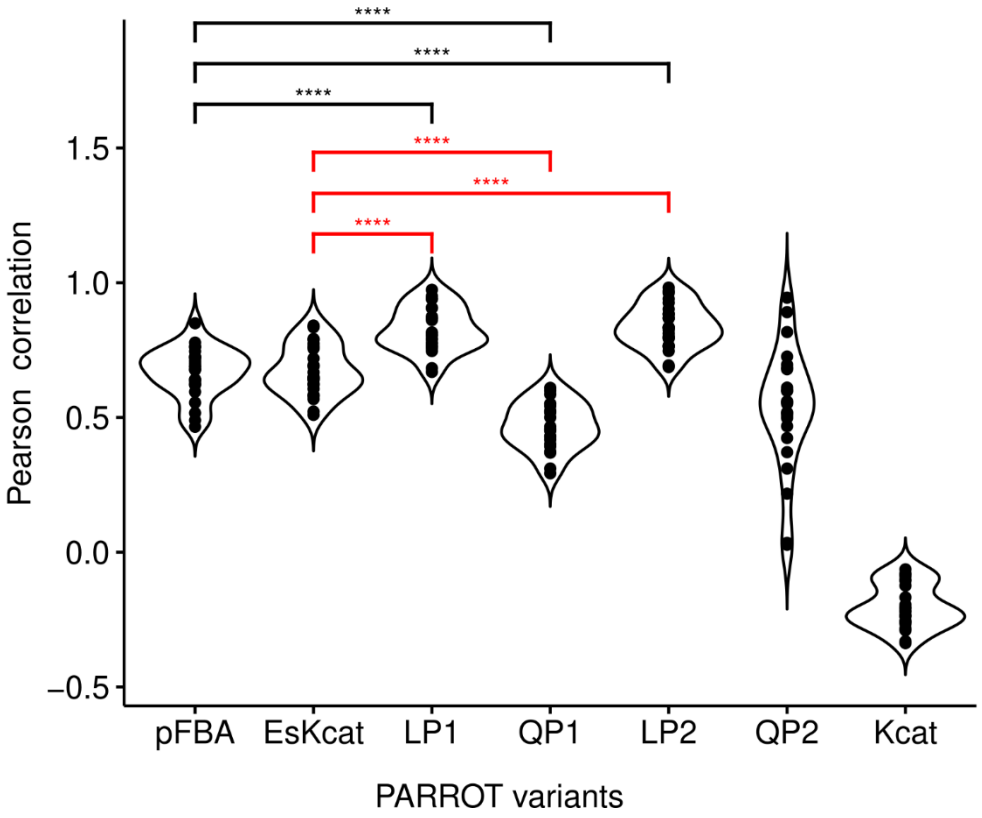

Fig. S3. Pearson correlation calculated between predicted enzyme distribution and the baseline obtained from minimizing the 2-norm of the experimental enzyme usage distribution, in *E. coli*. All values were log10-transformed prior to comparisons. A pairwise Wilcoxon rank sum assesses the statistical significance: **** p-value < 0.000005, * p-value < 0.03. Black significance bar indicates comparisons to pFBA. Red significance bar indicates comparisons to EsKcat.

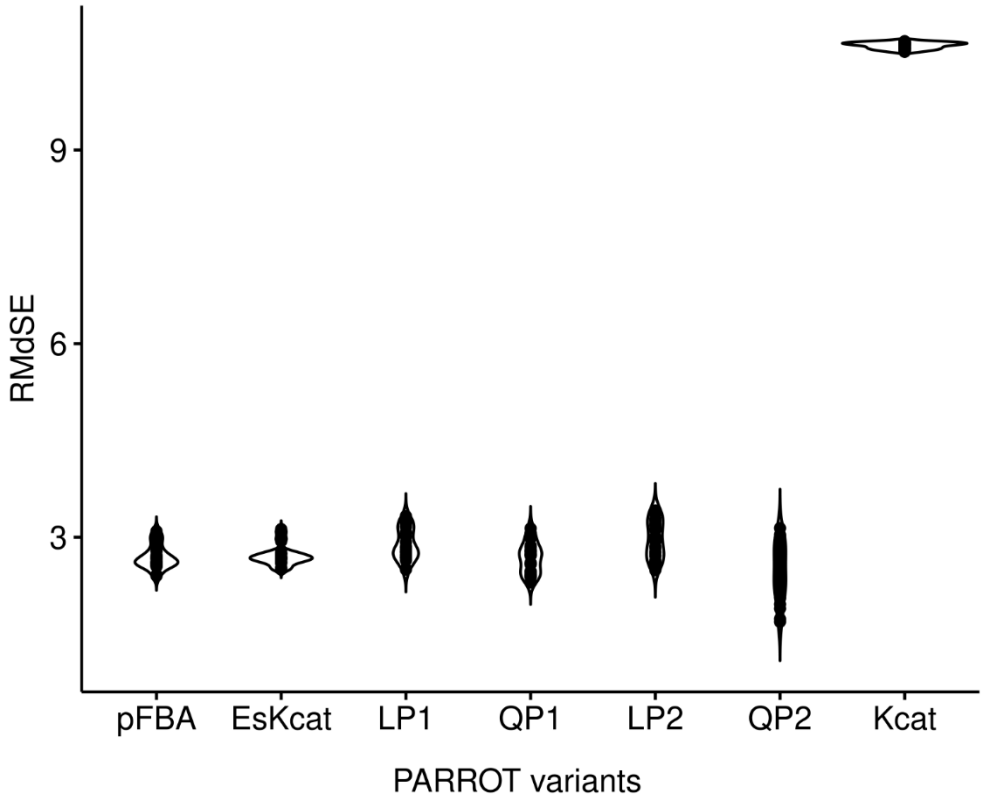

Fig. S4. Assessment of model performance based on the root median squared error (RMdSE). The minimization of the second norm of the experimental enzyme usage distribution in *E. coli* was used. All values were log10-transformed prior to comparisons.

Supplementary tables

Table S1 – Experimental proteomics measurements used for yeast

Condition	Usage by PARROT	Reference
Lahtvee2017_REF	Reference	Lahtvee et al. (2017)
Lahtvee2017_EtOH20	Alternative	
Lahtvee2017_EtOH40		
Lahtvee2017_EtOH60		
Lahtvee2017_Osmo02		
Lahtvee2017_Osmo04		
Lahtvee2017_Osmo06		
Yu2020_Clim	Reference	Yu et al. (2020)
Yu2020_CN30	Alternative	
Yu2020_CN50		
Yu2020_CN115		
Yu2021_std_010	Reference	Yu et al. (2021)
Yu2021_N30_005	Alternative	
Yu2021_N30_010		
Yu2021_N30_013		
Yu2021_N30_018		
Yu2021_N30_030		
Yu2021_N30_035		
Yu2021_Gln_glc1	Reference	
Yu2021_Gln_glc2	Alternative	
Yu2021_Gln_N30		
Yu2021_Phe_std	Reference	

Yu2021_Phe_N30	Alternative
Yu2021_Ile_std	Reference
Yu2021_Ile_N30	Alternative

Table S2 – Experimental proteomics measurements used for *Escherichia coli*

Condition	Usage by PARROT	Reference
GLYC_BATCH_mu=0.47_S	Reference	Schmidt et al. (2016)
ACE_BATCH_mu=0.3_S	Alternative	
GAM_BATCH_mu=0.46_S		
GLC_BATCH_mu=0.58_S		
MAN_BATCH_mu=0.47_S		
PYR_BATCH_mu=0.4_S		
XYL_BATCH_mu=0.55_S		
GLC_CHEM_mu=0.12_S	Reference	
GLC_CHEM_mu=0.20_S	Alternative	
GLC_CHEM_mu=0.35_S		
GLC_CHEM_mu=0.50_S		
GLC_CHEM_mu=0.11_V	Reference	Valgepea et al. (2013)
GLC_CHEM_mu=0.21_V	Alternative	
GLC_CHEM_mu=0.31_V		
GLC_CHEM_mu=0.40_V		
GLC_CHEM_mu=0.49_V		
GLC_CHEM_mu=0.21_P	Reference	Peebo et al. (2015)
GLC_CHEM_mu=0.22_P	Alternative	
GLC_CHEM_mu=0.26_P		
GLC_CHEM_mu=0.31_P		
GLC_CHEM_mu=0.36_P		
GLC_CHEM_mu=0.41_P		
GLC_CHEM_mu=0.46_P		
GLC_CHEM_mu=0.51_P		

APPENDIX C – SUPPLEMENTARY MATERIAL FOR CHAPTER 3.3

Supplementary material as available in the publication.

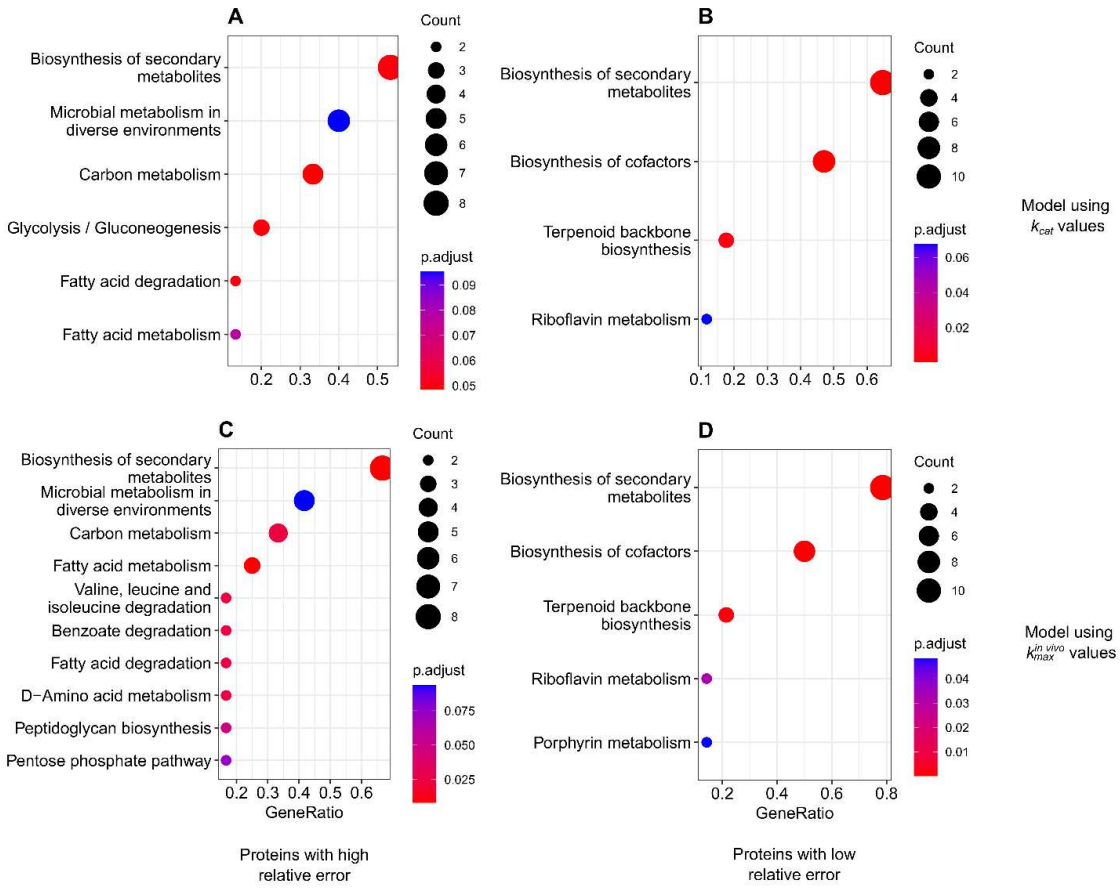

Figure S1. Enriched GO terms for proteins in *E. coli* with high and low relative error for the protein reserve ratio. Count represents the number of proteins assigned the GO term, and GeneRatio denotes the ratio between counts and the sample size. (A) Proteins with high error, model with *in vitro* $\boldsymbol{k_{cat}}$ values, (B) Proteins with low error, model with *in vitro* $\boldsymbol{k_{cat}}$ values (C) Proteins with high error, model with $\boldsymbol{k_{max}^{vivo}}$ values, (D) Proteins with low error, model with $\boldsymbol{k_{max}^{vivo}}$ values.

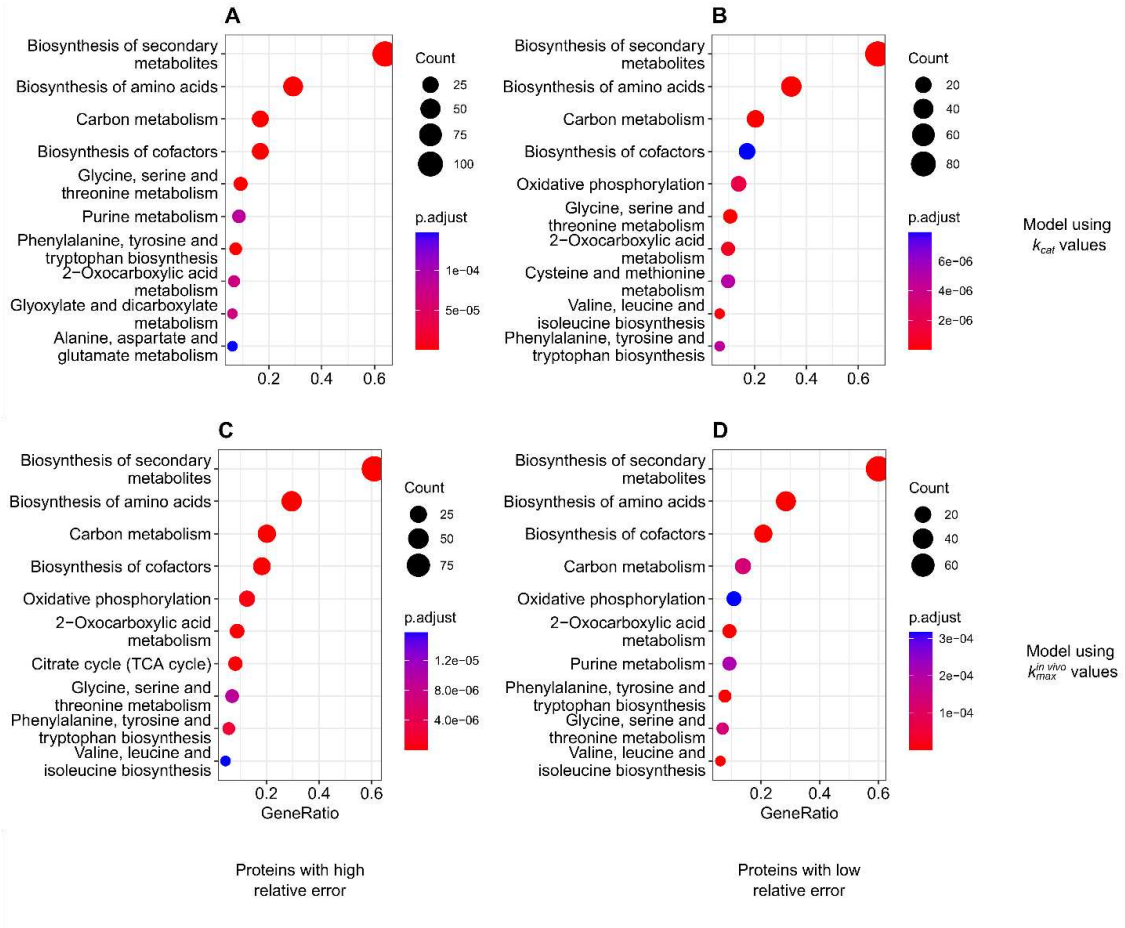

Figure S2. Enriched GO terms for proteins in *S. cerevisiae* with high and low relative error for the protein reserve ratio. Count represents the number of proteins assigned the GO term, and GeneRatio denotes the ratio between counts and the sample size. (A) Proteins with high error, model with *in vitro* k_{cat} values, (B) Proteins with low error, model with *in vitro* k_{cat} values (C) Proteins with high error, model with *in vivo* k_{max}^{vivo} values, (D) Proteins with low error, model with *in vivo* k_{max}^{vivo} values.

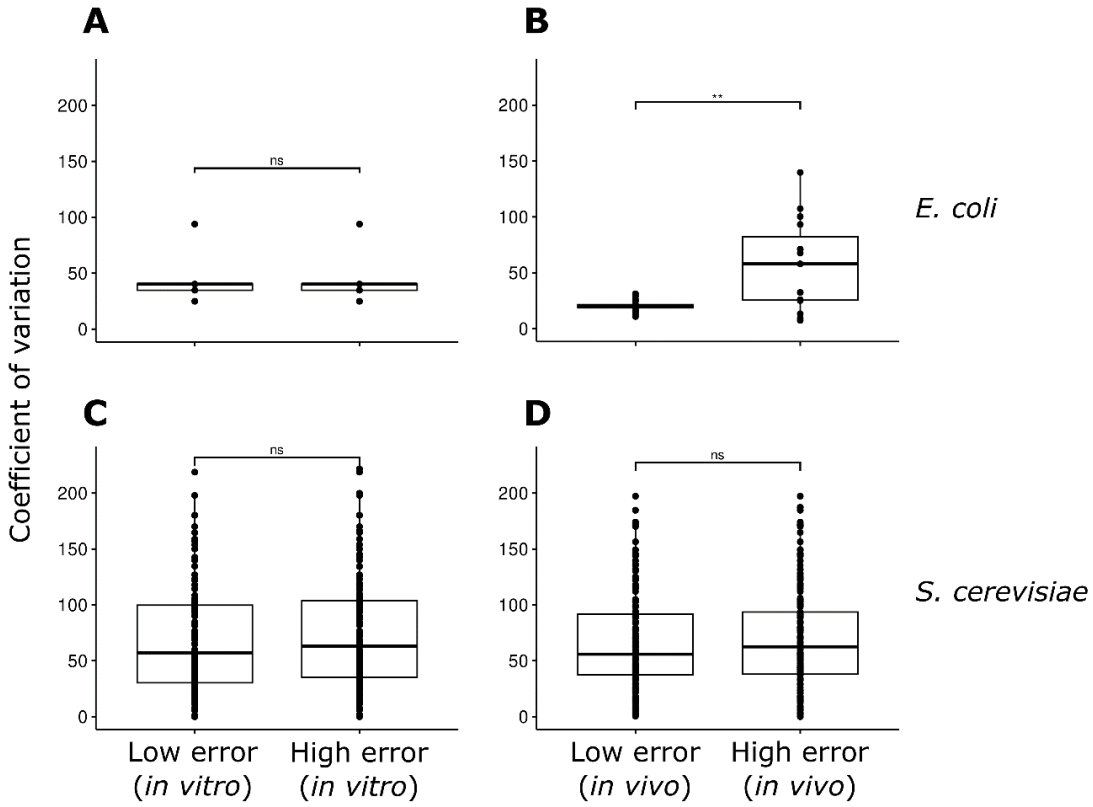

Figure S3. Comparison of the coefficient of variation between protein reserve ratios with low (≤25%) or high (≥100%) relative errors. *In vitro* refers to models using *in vitro* k_{cat} values, and *in vivo* refers to models using k_{max}^{vivo} values. (A) Comparison between *E. coli* models using *in vitro* k_{cat} values, (B) *E. coli* models using k_{max}^{vivo} values, (C) *S. cerevisiae* models using *in vitro* k_{cat} values, (D) *S. cerevisiae* models using k_{max}^{vivo} values. A pairwise Wilcoxon rank sum assesses the statistical significance: ** p-value < 0.01.

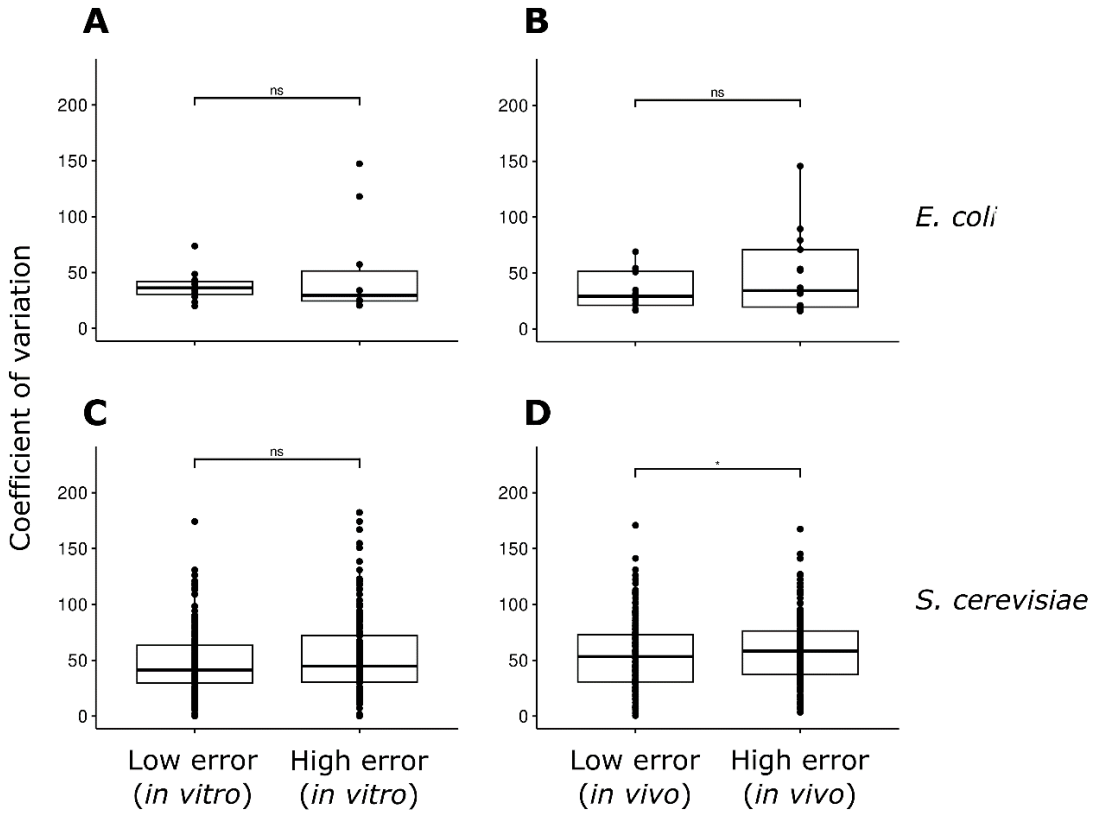

Figure S4. Comparison of the coefficient of variation between recalculated $E_s^{exp\prime}$ values with low (≤25%) or high (≥100%) relative errors. *In vitro* refers to models using *in vitro* k_{cat} values, and *in vivo* refers to models using k_{max}^{vivo} values. (A) Comparison between *E. coli* models using *in vitro* k_{cat} values, (B) *E. coli* models using k_{max}^{vivo} values, (C) *S. cerevisiae* models using *in vitro* k_{cat} values, (D) *S. cerevisiae* models using k_{max}^{vivo} values. A pairwise Wilcoxon rank sum assesses the statistical significance: * p-value < 0.05.

Table S1. Features included in the constructed data sets, for both *E. coli* and *S. cerevisiae*, using predicted $E_{s,i}$ values from models integrated with either k_{cat} or k_{max}^{vivo} values.

Features	**Type**	**Reference**
Predicted enzyme usage $E_{s,i}$	Condition-dependent	(Sánchez *et al.*, 2017; Domenzain *et al.*, 2022)
Metabolic flux	Condition-dependent	(Savinell and Palsson, 1992a, 1992b)
Information theory-based codon usage bias (iCUB)	Static	(Liu *et al.*, 2018)
tRNA adaptation index (tAI)	Static	(Reis *et al.*, 2004)
Codon adaptation index (CAI)	Static	(Sharp and Li, 1987)
Codon bias index (CBI)	Static	(Bennetzens and Hall, 1981)
Frequency of optimal codons (Fop)	Static	(Ikemura, 1981)
Effective number of codons (ENC)	Static	(Wright, 1990)
ENC alternative implementation (ENC')	Static	(Novembre, 2002)
G+C content of gene	Static	(Peden, 2000)
G+C of 3rd codon position	Static	(Peden, 2000)
Base composition at silent sites	Static	(Peden, 2000)
Hydropathicity of protein	Static	(Peden, 2000)
Aromaticity of protein	Static	(Peden, 2000)

B measure of codon bias	Static	(Karlin *et al.*, 2001)
E measure of expression	Static	(Karlin and Mrázek, 2000)
Maximum likelihood codon bias (MCB)	Static	(Urrutia and Hurst, 2001)
Measure independent of length and composition (MILC)	Static	(Supek and Vlahoviček, 2005)
MILC-based expression level predictor (MELP)	Static	(Supek and Vlahoviček, 2005)
Synonymous codon usage orderliness (SCUO)	Static	(Wan *et al.*, 2004)
Gene codon bias (GCB)	Static	(Merkl, 2003)
Evolutionary selection pressure on nucleotide biosynthetic cost (Sc)	Static	(Seward and Kelly, 2018, 2016)
Evolutionary selection pressure on gene translation efficiency (St)	Static	(Seward and Kelly, 2018, 2016)

Table S2. Scikit-Learn (Pedregosa *et al.*, 2011) functions employed in each pipeline optimized by TPOT. Abbreviations: SGDRegressor - Stochastic Gradient Descent regressor; XGBRegressor - eXtreme Gradient Boosting regressor; LassoLarsCV – Cross-validated Lasso using the least-angle regression algorithm.

***E. coli* (k_{cat})**	***E. coli* (k_{max}^{vivo})**	**Yeast (k_{cat})**	**Yeast (k_{max}^{vivo})**
Gradient Boosting Regressor	Function Transformer	RidgeCV	Function Transformer
SGD Regressor	Gradient Boosting Regressor	Normalizer (L1 norm)	PolynomialFeatures
Robust Scaler	LassoLarsCV	SGDRegressor	KneighborsRegressor
XGB Regressor	KNeighborsRegressor	XGBRegressor	ExtraTreesRegressor
	XGBRegressor		

Table S3. List of *E. coli* proteins with the highest relative error for the predicted protein reserve ratios, from models using either *in vitro* $\boldsymbol{k_{cat}}$ or $\boldsymbol{k_{max}^{vivo}}$ values.

Protein name	**EC number**	**Data set**
Triosephosphate isomerase	5.3.1.1	k_{cat}
Malate dehydrogenase	1.1.1.37	k_{cat}
3-ketoacyl-CoA thiolase	2.3.1.16	k_{cat}
Fatty acid oxidation complex subunit alpha	4.2.1.17	k_{cat}
Guanyl-specific ribonuclease	4.6.1.24	k_{cat}
2,3-bisphosphoglycerate-dependent phosphoglycerate mutase	5.4.2.11	k_{max}^{vivo}
UDP-N-acetylmuramoylalanine-D-glutamate ligase	6.3.2.9	k_{max}^{vivo}
UDP-N-acetylmuramate-L-alanine ligase	6.3.2.8	k_{max}^{vivo}
3-oxoacyl-[acyl-carrier-protein] synthase 1	2.3.1.293	k_{max}^{vivo}
Acyl carrier protein	2.3.1.40	k_{max}^{vivo}

Table S4. List of S*. cerevisiae* proteins with the highest relative error for the predicted protein reserve ratios, from models using either *in vitro* $\boldsymbol{k_{cat}}$ or $\boldsymbol{k_{max}^{vivo}}$ values.

Protein name	**EC number**	**Data set**
4-aminobutyrate aminotransferase	2.6.1.19	k_{cat}
Bifunctional purine biosynthesis protein	2.1.2.3	k_{cat}
Mitochondrial glycine dehydrogenase	1.4.4.2	k_{cat}
Phosphoglucomutase 2	5.4.2.2	k_{cat} and k_{max}^{vivo}
Delta-aminolevulinic acid dehydratase	4.2.1.24	k_{cat}
Mitochondrial inorganic pyrophosphatase	3.6.1.1	k_{max}^{vivo}
Porphobilinogen deaminase	2.5.1.61	k_{max}^{vivo}
Cytochrome b-c1 complex subunit 8	7.1.1.8	k_{max}^{vivo}
Isocitrate dehydrogenase [NADP]	1.1.1.42	k_{max}^{vivo}

Table S5. Pearson correlations between measured (E_s^{exp}) and predicted ($E_s^{exp'}$) protein abundances for each growth condition in *E. coli* using k_{cat} values.

Growth condition	Pearson correlation	p-value	Number of proteins
ACE_BATCH_mu=0.3_S	0.997	1,74E+02	12
GAM_BATCH_mu=0.46_S	0.891	1,90E+04	28
GLC_BATCH_mu=0.58_S	0.998	6,69E-18	26
GLC_CHEM_mu=0.11_V	0.979	9,29E-11	34
GLC_CHEM_mu=0.12_S	0.979	6,88E+00	20
GLC_CHEM_mu=0.20_S	0.997	9,28E-18	28
GLC_CHEM_mu=0.21_P	0.993	2,88E-18	34
GLC_CHEM_mu=0.21_V	0.995	2,73E-14	28
GLC_CHEM_mu=0.22_P	0.998	4,28E-31	37
GLC_CHEM_mu=0.26_P	0.986	7,40E-08	28
GLC_CHEM_mu=0.31_P	0.988	2,58E+09	7
GLC_CHEM_mu=0.31_V	0.988	1,20E+06	11
GLC_CHEM_mu=0.35_S	0.998	1,67E-05	17
GLC_CHEM_mu=0.36_P	0.995	1,04E+00	15
GLC_CHEM_mu=0.40_V	0.998	5,24E-14	24
GLC_CHEM_mu=0.41_P	0.995	1,97E+00	14
GLC_CHEM_mu=0.46_P	0.998	2,54E-06	17
GLC_CHEM_mu=0.49_V	0.954	1,44E+08	12
GLC_CHEM_mu=0.50_S	0.992	3,07E+01	15
GLC_CHEM_mu=0.51_P	0.997	1,62E-09	21
GLYC_BATCH_mu=0.47_S	0.991	2,44E-04	21
MAN_BATCH_mu=0.47_S	0.994	3,64E-13	28
PYR_BATCH_mu=0.4_S	0.997	4,50E-18	29
XYL_BATCH_mu=0.55_S	0.999	6,39E-21	25

Table S6. Pearson correlations between measured (E_s^{exp}) and predicted ($E_s^{exp'}$) protein abundances for each growth condition in *E. coli* using k_{max}^{vivo} values.

Growth condition	Pearson correlation	p-value	Number of proteins
ACE_BATCH_mu=0.3_S	0.997	3,43E-18	28
GAM_BATCH_mu=0.46_S	0.995	4,64E-07	20
GLC_BATCH_mu=0.58_S	0.996	1,18E-07	21
GLC_CHEM_mu=0.11_V	0.991	5,70E-10	26
GLC_CHEM_mu=0.12_S	0.990	2,29E+01	16
GLC_CHEM_mu=0.20_S	0.989	3,32E+01	16
GLC_CHEM_mu=0.21_P	0.998	1,06E-21	29
GLC_CHEM_mu=0.21_V	0.996	3,66E-14	26
GLC_CHEM_mu=0.22_P	0.991	4,18E-07	24
GLC_CHEM_mu=0.26_P	0.998	2,27E-03	14
GLC_CHEM_mu=0.31_P	0.995	5,68E-08	22
GLC_CHEM_mu=0.31_V	0.999	2,71E-02	11
GLC_CHEM_mu=0.35_S	0.996	3,95E-13	26
GLC_CHEM_mu=0.36_P	0.989	2,68E-08	27
GLC_CHEM_mu=0.40_V	0.998	3,25E-14	22
GLC_CHEM_mu=0.41_P	0.983	1,48E-04	25
GLC_CHEM_mu=0.46_P	0.999	7,91E-09	17
GLC_CHEM_mu=0.49_V	0.988	3,12E+02	14
GLC_CHEM_mu=0.50_S	0.996	5,62E-16	27
GLC_CHEM_mu=0.51_P	0.992	7,83E-04	20
GLYC_BATCH_mu=0.47_S	0.998	3,64E-34	36
MAN_BATCH_mu=0.47_S	0.995	3,53E-05	18
PYR_BATCH_mu=0.4_S	0.997	2,68E-11	22
XYL_BATCH_mu=0.55_S	0.994	4,01E-03	18

Table S7. Pearson correlations between measured (E_s^{exp}) and predicted ($E_s^{exp'}$) protein abundances for each growth condition in *S. cerevisiae* using k_{cat} values.

Growth condition	Pearson correlation	p-value	Number of proteins
Lahtvee2017_EtOH20	0.465	0.006	33
Lahtvee2017_EtOH40	0.328	0.044	38
Lahtvee2017_EtOH60	0.295	0.080	36
Lahtvee2017_Osmo02	0.150	0.330	44
Lahtvee2017_Osmo04	0.020	0.918	27
Lahtvee2017_Osmo06	0.442	0.031	24
Lahtvee2017_REF	0.074	0.652	39
Yu2020_Clim	0.253	0.101	43
Yu2020_CN115	0.717	6,06E+06	43
Yu2020_CN30	0.540	0.002	29
Yu2020_CN50	0.706	1,22E+06	43
Yu2021_Gln_glc1	0.739	2,16E+06	42
Yu2021_Gln_glc2	0.489	0.003	33
Yu2021_Gln_N30	0.559	0.0004	35
Yu2021_Ile_N30	0.785	5,18E+05	38
Yu2021_Ile_std	0.697	2,02E+07	43
Yu2021_N30_005	0.535	0.0001	44
Yu2021_N30_010	0.578	0.0003	34
Yu2021_N30_013	0.487	0.0005	46
Yu2021_N30_018	0.336	0.0387	38
Yu2021_N30_030	0.501	0.0005	44
Yu2021_Phe_N30	0.672	3,71E+07	38
Yu2021_Phe_std	0.700	2,82E+08	35
Yu2021_std_010	0.531	0.0008	36

Table S8. Pearson correlations between measured (E_s^{exp}) and predicted ($E_s^{exp'}$) protein abundances for each growth condition in *S. cerevisiae* using k_{max}^{vivo} values.

Growth condition	Pearson correlation	p-value	Number of proteins
Lahtvee2017_EtOH20	0.150	0.366	38
Lahtvee2017_EtOH40	0.207	0.204	39
Lahtvee2017_EtOH60	0.430	0.012	33
Lahtvee2017_Osmo02	0.348	0.034	37
Lahtvee2017_Osmo04	0.456	0.012	29
Lahtvee2017_Osmo06	-0.107	0.539	35
Lahtvee2017_REF	0.434	0.023	27
Yu2020_Clim	0.633	5,82E+09	34
Yu2020_CN115	0.591	0.0002	33
Yu2020_CN30	0.730	4,00E+06	42
Yu2020_CN50	0.629	2,99E+09	37
Yu2021_Gln_glc1	0.678	5,38E+08	36
Yu2021_Gln_glc2	0.590	4,84E+09	41
Yu2021_Gln_N30	0.524	0.0001	48
Yu2021_Ile_N30	0.859	3,03E+01	42
Yu2021_Ile_std	0.845	1,72E+04	35
Yu2021_N30_005	0.630	3,46E+08	45
Yu2021_N30_010	0.564	0.0009	31
Yu2021_N30_013	0.496	0.001	40
Yu2021_N30_018	0.495	0.0006	44
Yu2021_N30_030	0.443	0.002	46
Yu2021_Phe_N30	0.699	5,12E+07	40
Yu2021_Phe_std	0.612	0.0001	33
Yu2021_std_010	0.370	0.017	41

Table S9. List of *E. coli* proteins with the highest relative error for the recalculated $\boldsymbol{E_s^{exp\prime}}$ values, coming from models using either $\boldsymbol{k_{cat}}$ or $\boldsymbol{k_{max}^{vivo}}$ values.

Protein name	**EC number**	**Data set**
Molybdate-binding protein	7.3.2.5	k_{cat}
3-phosphoshikimate 1-carboxyvinyltransferase	2.5.1.19	k_{cat}
Fatty acid oxidation complex subunit alpha	4.2.1.17	k_{cat}
6-phosphogluconate dehydrogenase	1.1.1.44	k_{cat} and k_{max}^{vivo}
Glutamate racemase	5.1.1.3	k_{cat}
Flavodoxin 1	1.18.1.2	k_{max}^{vivo}
Adenylate kinase	2.7.4.3	k_{max}^{vivo}
Ribose-phosphate pyrophosphokinase	2.7.6.1	k_{max}^{vivo}
3-oxoacyl-[acyl-carrier-protein] synthase 1	2.3.1.293	k_{max}^{vivo}

Table S10. List of *S. cerevisiae* proteins with the highest relative error for the recalculated $\boldsymbol{E_s^{exp\prime}}$ values, coming from models using either $\boldsymbol{k_{cat}}$ or $\boldsymbol{k_{max}^{vivo}}$ values.

Protein name	**EC number**	**Data set**
Pyruvate decarboxylase isozyme 1	4.1.1.1	k_{cat}
Mitochondrial glycine dehydrogenase	1.4.4.2	k_{cat} and k_{max}^{vivo}
Asparagine synthetase 1	6.3.5.4	k_{cat}
3-hydroxy-3-methylglutaryl-coenzyme A reductase 2	1.1.1.34	k_{cat}
Pyruvate dehydrogenase complex protein X component	1.2.4.1	k_{cat} and k_{max}^{vivo}
Mitochondrial inorganic pyrophosphatase	3.6.1.1	k_{max}^{vivo}
NADPH-cytochrome P450 reductase	1.6.2.4	k_{max}^{vivo}
Ornithine carbamoyltransferase	2.1.3.3	k_{max}^{vivo}

Table S11. Flux variability analysis for central metabolic pathways and enzyme usage pseudo-reactions. Median flux ratio was calculated by dividing the median maximum flux by the median minimum flux across all growth conditions.

Species	**Reactions**	**Dataset**	**Median flux ratio**
E. coli	Central metabolic pathways	k_{cat}	64,88342
		k_{max}^{vivo}	52,20936
	Enzyme usage pseudo-reactions	k_{cat}	1,000059
		k_{max}^{vivo}	1,000048
S. cerevisiae	Central metabolic pathways	k_{cat}	16,29839
		k_{max}^{vivo}	18,62662
	Enzyme usage pseudo-reactions	k_{cat}	1,000092
		k_{max}^{vivo}	1,000089

SI References

Bennetzens,J.L. and Hall,B.D. (1981) Codon Selection in Yeast. *J Biol Chem*, **257**, 3026–3031.

Domenzain,I. *et al.* (2022) Reconstruction of a catalogue of genome-scale metabolic models with enzymatic constraints using GECKO 2.0. *Nat Commun*, **13**, 1–13.

Ikemura,T. (1981) Correlation between the abundance of Escherichia coli transfer RNAs and the occurrence of the respective codons in its protein genes: A proposal for a synonymous codon choice that is optimal for the E. coli translational system. *J Mol Biol*, **151**, 389–409.

Karlin,S. *et al.* (2001) Characterizations of Highly Expressed Genes of Four Fast-Growing Bacteria. *Society*, **183**, 5025–5040.

Karlin,S. and Mrázek,J. (2000) Predicted highly expressed genes of diverse prokaryotic genomes. *J Bacteriol*, **182**, 5238–50.

Liu,S.S. *et al.* (2018) A novel framework for evaluating the performance of codon usage bias metrics. *J R Soc Interface*, **15**, 20170667.

Merkl,R. (2003) A Survey of Codon and Amino Acid Frequency Bias in Microbial Genomes Focusing on Translational Efficiency. *J Mol Evol*, **57**, 453–466.

Novembre,J.A. (2002) Accounting for Background Nucleotide Composition When Measuring Codon Usage Bias. *Mol Biol Evol*, **19**, 1390–1394.

Peden,J.F. (2000) Analysis of Codon Usage.

Pedregosa,F. *et al.* (2011) Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*, **12**, 2825–2830.

Reis,M. d. *et al.* (2004) Solving the riddle of codon usage preferences: a test for translational selection. *Nucleic Acids Res*, **32**, 5036–5044.

Sánchez,B.J. *et al.* (2017) Improving the phenotype predictions of a yeast genome-scale metabolic model by incorporating enzymatic constraints. *Mol Syst Biol*, **13**, 935.

Savinell,J.M. and Palsson,B.O. (1992a) Optimal selection of metabolic fluxes for in vivo measurement. I. Development of mathematical methods. *J Theor Biol*, **155**, 201–214.

Savinell,J.M. and Palsson,B.O. (1992b) Optimal selection of metabolic fluxes for in vivo measurement. II. Application to Escherichia coli and hybridoma cell metabolism. *J Theor Biol*, **155**, 215–242.

Seward,E.A. and Kelly,S. (2016) Dietary nitrogen alters codon bias and genome composition in parasitic microorganisms. *Genome Biol*, **17**, 226.

Seward,E.A. and Kelly,S. (2018) Selection-driven cost-efficiency optimization of transcripts modulates gene evolutionary rate in bacteria. *Genome Biol*, **19**, 1–11.

Sharp,P.M. and Li,W.-H. (1987) The codon adaptation index-a measure of directional synonymous codon usage bias, and its potential applications. *Nucleic Acids Res*, **15**, 1281–1295.

Supek,F. and Vlahoviček,K. (2005) Comparison of codon usage measures and their applicability in prediction of microbial gene expressivity. *BMC Bioinformatics*, **6**, 182.

Urrutia,A.O. and Hurst,L.D. (2001) Codon usage bias covaries with expression breadth and the rate of synonymous evolution in humans, but this is not evidence for selection. *Genetics*, **159**, 1191–9.

Wan,X.-F. *et al.* (2004) Quantitative relationship between synonymous codon usage bias and GC composition across unicellular genomes. *BMC Evol Biol*, **4**, 19.

Wright,F. (1990) The ‘effective number of codons’ used in a gene. *Gene*, **87**, 23–29.

APPENDIX D – SUPPLEMENTARY MATERIAL FOR CHAPTER 3.4

Original text not yet submitted to a journal or preprint repository.

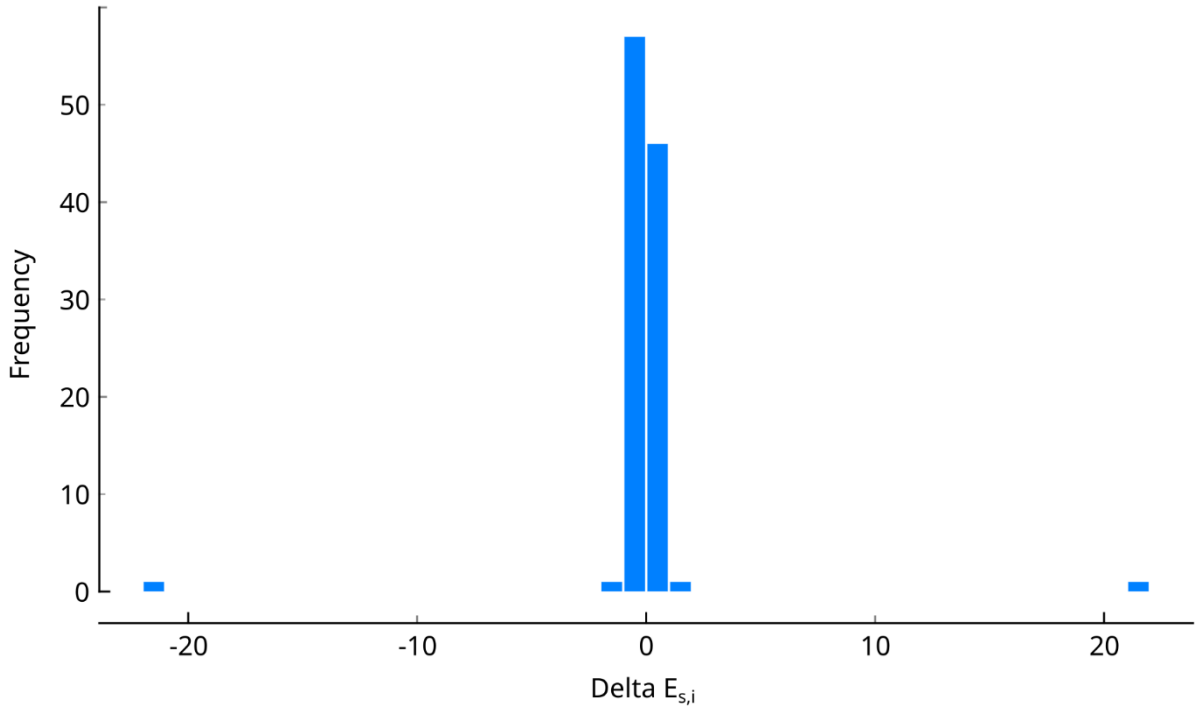

Figure S1. **Changes in enzyme subpool usage following the knockout of a single main reaction.** We calculated delta by subtracting the enzyme subpool usage of the wildtype from the enzyme subpool usage of the mutant solutions.

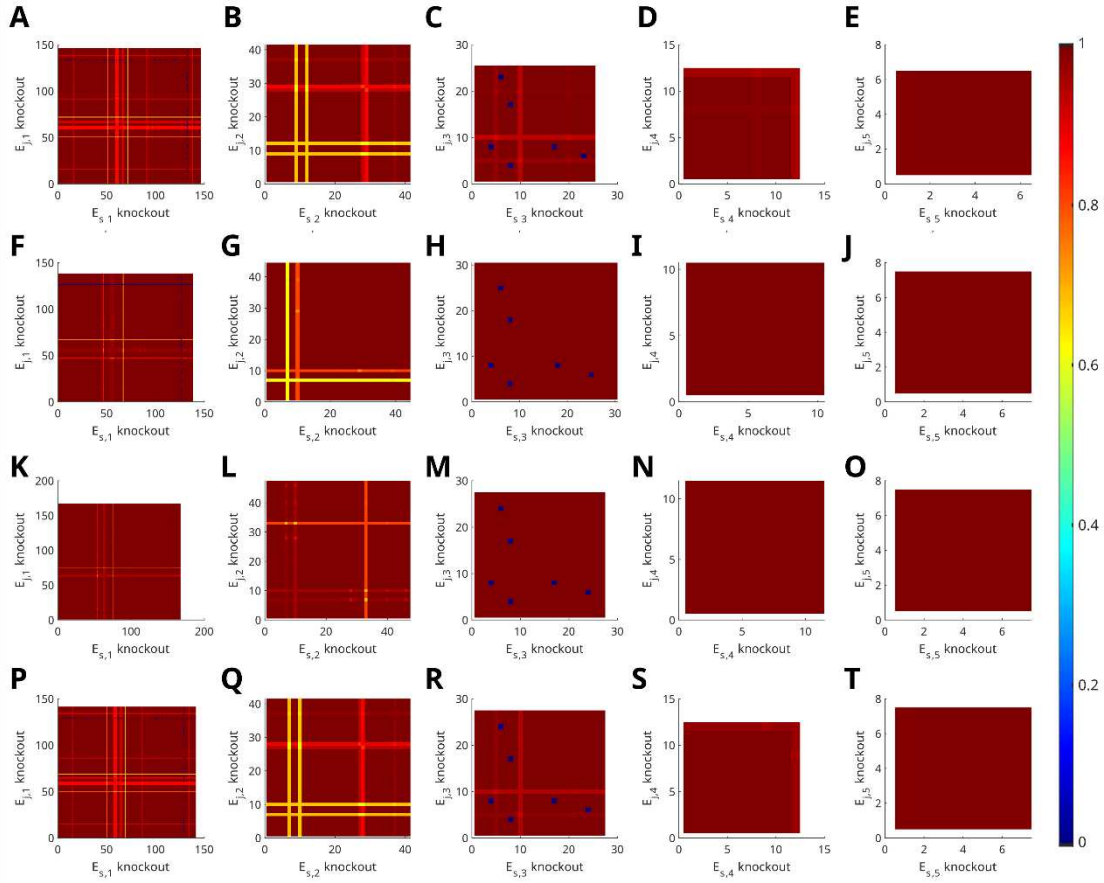

Figure S2. **Impact on growth after double knockouts considering alternative carbon sources. A)** Impact on growth after a double knockout of first subpool ($E_{s,1}$) using arabinose as carbon source. **B)** Second subpool ($E_{s,2}$). **C)** Third subpool ($E_{s,3}$). **D)** Fourth subpool ($E_{s,4}$). **E)** Fifth subpool ($E_{s,5}$). **F)** Impact on growth after a double knockout of first subpool ($E_{s,1}$) using fructose as carbon source. **G)** Second subpool ($E_{s,2}$). **H)** Third subpool ($E_{s,3}$). **I)** Fourth subpool ($E_{s,4}$). **J)** Fifth subpool ($E_{s,5}$). **K)** Impact on growth after a double knockout of first subpool ($E_{s,1}$) using fucose as carbon source. **L)** Second subpool ($E_{s,2}$). **M)** Third subpool ($E_{s,3}$). **N)** Fourth subpool ($E_{s,4}$). **O)** Fifth subpool ($E_{s,5}$). **P)** Impact on growth after a double knockout of first subpool ($E_{s,1}$) using xylose as carbon source. **Q)** Second subpool ($E_{s,2}$). **R)** Third subpool ($E_{s,3}$). **S)** Fourth subpool ($E_{s,4}$). **T)** Fifth subpool ($E_{s,5}$).

Description of Additional Supplementary Files

The Supplementary Data is available as Excel spreadsheets in the GitHub repository: https://github.com/mauricioamf/CORAL.

File Name: Supplementary Data 1
Description: List of subpools and their corresponding percentage of the enzyme pool.

File Name: Supplementary Data 2
Description: List of knockout pairs and their corresponding results (subpool usage before and after knockout, growth ratio before and after knockout).