MODEL-GUIDED STRAIN ENGINEERING

Bridges between Design and Learning in Bioprocess Development

SARA MORENO PAZ

Propositions

1. Bioprocesses will only be optimal when genetic and environmental factors are simultaneously considered.
   (this thesis)

2. The optimization of bioprocesses and strains unveils fundamental biological principles.
   (this thesis)

3. The sustainability of the bio-based economy is frequently presumed rather than demonstrated.

4. Scientists unconsciously underestimate the potential of simple ideas.

5. The pressure for publishing hinders high-risk research and thus scientific breakthroughs.

6. Embracing and acknowledging one's vulnerabilities is a sign of self-awareness and strength that fosters open communication.

7. The advent of Artificial Intelligence will improve the work-life balance.

Propositions belonging to the thesis, entitled

"Model-guided strain engineering: bridges between design and learning in bioprocess development"

Sara Moreno Paz

Wageningen, 31 May 2024

Model-Guided Strain Engineering: Bridges between Design and Learning in Bioprocess Development

Sara Moreno Paz

Thesis committee

Promotors

Prof. Dr Vitor A.P. Martins dos Santos

Personal Chair, Bioprocess Engineering

Wageningen University & Research

Prof. Dr María Suárez Diez

Professor of Systems and Synthetic Biology

Wageningen University & Research

Co-promotor

Dr Joep Schmitz

R&D manager Bioinformatics and Modeling

dsm-firmenich, Delft

Other members

Prof. Dr D. Z. Machado de Sousa, Wageningen University & Research

Prof. Dr P. Carbonell Cortés, Universitat Politècnica de València, Spain

Prof. Dr I. Rocha, Universidade Nova de Lisboa, Portugal

Dr P. Berends, Johnson & Johnson Innovative Medicines, Leiden

This research was conducted under the auspices of VLAG Graduate School (Biobased, Biomolecular, Chemical, Food, and Nutrition Sciences).

Model-Guided Strain Engineering:

Bridges between Design and Learning in Bioprocess Development

Sara Moreno Paz

Thesis

submitted in fulfilment of the requirements for the degree of doctor

at Wageningen University

by the authority of the Rector Magnificus,

Prof. Dr C. Kroeze,

in the presence of the

Thesis Committee appointed by the Academic Board

to be defended in public

on Friday 31 May 2024

at 1.30 p.m. in the Omnia Auditorium.

Sara Moreno Paz

Model-Guided Strain Engineering:

Bridges between Design and Learning in Bioprocess Development,

289 pages.

PhD thesis, Wageningen University, Wageningen, the Netherlands (2024)

With references, with summary in English

ISBN: 978-94-6469-927-2
DOI: 10.18174/656101

Contents

CHAPTER 1

General Introduction

Sara Moreno Paz

Preface

In 2013 I started a Bachelor's degree in Biotechnology followed by a Master's and, finally, in 2020, I started this PhD. Over this period, my perspective on the role of biotechnology in society evolved. As I learned, I recognized the immense potential of biotechnology but, also, its inherent limitations. Like me, many people acknowledge the power of biotechnology to address global challenges, from pandemics to climate change. However, others still associate biotechnology with fear of the unknown, scientists that play with nature, and powerful corporations that benefit from it. Meeting biotechnology's promises while changing public perception is a difficult challenge to overcome. Yet, it serves as a compelling motivation to contribute, even in the slightest form, to the technical advancement of this discipline and to foster a broader understanding of biotechnology among the general public.

As implied by its title, this thesis delves into the use of mathematical models for the design of microbial cell factories and bioprocesses. However, before diving into these concepts, this introductory chapter provides a glimpse into the achievements, promises, and challenges of biotechnology in a comprehensible manner. Subsequently, Design-Build-Test-Learn cycles, one of the foundations of this thesis, are presented. This is followed by an exploration of the models applicable to the Design and Learn phases of these cycles. Finally, I outline the objectives and structure of the thesis.

A brief history of (microbial) biotechnology

Accomplishments of Biotechnology

Biotechnology harnesses the power of biology to create new services and products that improve the quality of our lives and the environment [1]. In essence, biotechnologists seek to understand nature, specifically living organisms, to create solutions to prevalent challenges. These organisms include microorganisms such as bacteria, yeast, or fungi, as well as algae, plants and animals. My focus within biotechnology has centered on microbial biotechnology, with the objective of understanding the functioning of microorganisms. These microorganisms can then be used and modified to produce valuable and diverse products, ranging from food and medicines to vitamins, aromas, preservatives, or plastics. Biotechnology is not new and, within the broad spectrum of biotechnological achievements, I have highlighted some accomplishments that I find particularly noteworthy.

Since Neolithic times, humankind has been using biotechnology for food production and we still use it to produce beer, wine, bread, cheese, or tempeh [2]. In all these processes microorganisms ferment sugars into alcohol and CO_2 or degrade complex molecules into simpler ones. Although we had used fermentation technology for millennia, it was not until the XIX century that the role of microorganisms in this process was understood [3]. Since then, we have developed sophisticated techniques that endow us with a growing ability to use these microorganisms to solve complex problems.

Some of the undisputed accomplishments of biotechnology lie in the medical area. What started with the accidental discovery of penicillin, led to the efficient production of microbial-based antibiotics, vaccines, and the production of recombinant proteins such as insulin or monoclonal antibodies [3, 4]. Individualized therapies such as the use of CAR T cells, that specifically target the immune response against cancer cells, continue to show the potential of biotechnology in this area [5]. In 2020, with the COVID-19 pandemic, the power of biotechnology to solve a global challenge by developing vaccines in record times became evident [6].

Biotechnology also contributes to the continuous increase in food production, necessary to meet the needs of the growing world population [7]. Crop domestication, which started 13,000 years ago, transformed wild plants into variants with accessible fruits and grains. Since then, plant biotechnology has achieved the development of pest-resistant crops or crop variants with better nutritional traits [8]. Besides, as an alternative to fertilizers, beneficial microorganisms are now used to enhance the quality of crops, their nutrient uptake, or their tolerance to stress [9].

Although less noticeable, biotechnology is present in many other aspects of our lives. At the beginning of the XX century, Weizmann established a large-scale butanol fermentation process by *Clostridia* that remained the main source of butanol until a more competitive petrochemical production was established in 1960. Since 1955 microbially produced amino acids have been available and biotechnological processes to produce bulk chemicals such as citric acid, an important food preservative, have been developed [3]. Flavors and fragrances such as nootkatone

or vanillin made by fermentation are also commercialized [3]. In 1956 microbial enzymes reached consumers when a detergent containing alcalase obtained from *Bacillus licheniformis* was commercialized and, since then, recombinant proteins have been standard ingredients of household detergents [3]. However, it was not until the first energy crisis in 1973, that biotechnology was explored as an alternative to a fossil fuel-based economy, and, at the beginning of the XXI century, it became a potential asset in the fight against climate change [3].

Biotechnology promises

In 2015 the European Commission presented its first action plan for a circular economy [10]. They defined circular economy as the minimization of waste generation and the maintenance of the value of products, materials, and resources for as long as possible [10]. This concept combines the requirement for sustainable development that ensures the care of the environment, with the support of economic growth and the creation of jobs and business opportunities [11]. The bio-economy, defined as the "production of renewable biological resources and the conversion of these resources and waste streams into value-added products, such as food, feed, bio-based products, and bioenergy", was considered one of the priority areas to achieve a circular economy [12]. Besides the European Union and the United States, bioeconomy policy strategies have expanded to highly industrialized countries, transition economies, and developing nations [1, 4, 13].

A successful circular economy is the biggest promise of biotechnology. Biorefineries have been envisioned as alternatives to oil refineries, where waste (including side streams like glycerol, lignocellulosic biomass, or CO_2) is up-cycled into high-value products using biotechnology and creating profit [11]. Biorefineries need to compete with traditional petrochemical processes that have been optimized for decades, while being strict regarding environmental impact [14, 15]. However, there are trade-offs between sustainability and economic growth and, while the former is assumed, the latter is prioritized [16]. Therefore, although using waste and producing bulk products contributes to the fight against climate change, more opportunities are currently envisioned in the generation of products with high economic value in sectors such as cosmetics, food additives, or pharmaceuticals using simpler sugars as substrates [2, 14, 17]. These processes assume positive effects on climate change and environmental aspects and exploit the value creation trait included in the circular and bio-economy definitions [16]. Yet, technological breakthroughs achieved in these industries can drive the development of biotechnology as a whole and have the potential to translate to more competitive and relevant sectors.

Moving towards a circular, bio-based economy is imperative, particularly in light of the recognized need to transition away from all fossil fuels, acknowledged during the 28th Conference of the Parties (COP28) [18]. However, achieving it requires both technological advances and policy changes [19]. For instance, changes in regulation that facilitate the use of "waste" as starting material for biotechnology are required to foster investments in biorefineries [19]. Similarly, the end of fossil fuel subsidies or explicit taxes on CO_2 emissions are a necessity to ensure the economic

success of bioprocesses [19]. Besides, social acceptance of industrial biotechnology is a crucial issue. Although the public is enthusiastic about the use of biotechnology for medical advances, acceptance of biotechnological applications that help to reduce greenhouse gas emissions is less common [20, 21]. Changing this perception requires improving societal trust in governments and treating citizens as key players during legislation [21].

In addition to climate change, biotechnology also promises solutions that contribute to meeting the sustainable development goals adopted by all United Nations Member States in 2015 [22]. For example, the development of alternative protein sources contributes to the end hunger goal, new biopharmaceuticals contribute to ensuring healthy lives, the development of wastewater treatments is required for water sanitation, the production of bio-based chemicals contributes to sustainable consumption, and the development of efficient crops, resistant to plagues, is required to preserve terrestrial ecosystems [23, 24].

Biotechnology challenges

So far, biotechnology has been defined in general terms. From now on, in this thesis, I will focus on industrial microbial biotechnology, also known as biomanufacturing, which uses microorganisms, usually referred to as cell factories, for the synthesis of chemical products [25] (Figure 1.1). Currently, about 200 out of the 70,000 chemicals commercially available are estimated to be produced by bioprocesses [14]. The tools employed to construct cell factories include synthetic biology and metabolic engineering. These terms are often used interchangeably and, while synthetic biology is the field of science that involves redesigning organisms for useful purposes by engineering them to have new abilities, metabolic engineering is the targeted genetic modification of cell factories to produce novel chemicals and/or improve product yield [15].

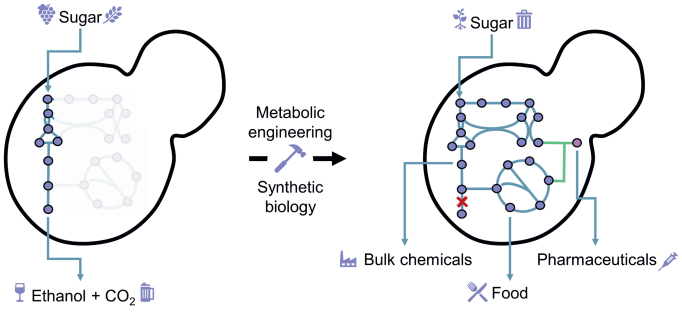


Figure 1.1: Microorganisms such as yeast can naturally convert sugars present in grapes or wheat to ethanol and CO_2 resulting in wine and beer production. This conversion is performed through a series of biochemical reactions called metabolism. Using metabolic engineering and synthetic biology we can convert yeasts into cell factories. These cell factories do not only uptake sugar but they could also use waste as substrate. By avoiding some biochemical reactions (red cross), activating pathways, or including new reactions (green lines), the cell factory can produce a variety of products such as bulk chemicals, food ingredients, or pharmaceuticals.

Compared with other technologies, the main challenge of industrial biotechnology resides in the lack of complete understanding of cell function [15]. When engineers build a machine, the function of each of its components is known and only relevant parts are included in the design. However, when biotechnologists design a cell factory, they generally have to adapt an already existing system (the cell) that has not evolved to fulfill the designer's purpose. This is done without complete knowledge of the function of each element in the cell, or how these elements interact with each other and with the desired new functionality. The performance of the cell factory is measured in terms of titer, rate, and yield. While the titer is the final concentration of the product, the yield refers to how efficiently the substrate is used, and the rate considers the time required for production [14]. Many factors can affect the performance of a cell factory including the selection of the host organism, the production pathway (length, burden, toxic intermediates, level of expression of genes), and the production conditions [14, 25]. Moreover, scaling bio-based processes is more challenging than scaling traditional chemical processes, as, when scaled, the microorganism does not necessarily behave as the laboratory-scale results would predict [26].

The lack of complete knowledge of cell metabolism and physiology prevents predictions on the behavior of a microorganism when genetic or environmental factors are perturbed [15]. Hence, bioprocess engineers have to go through intensive experimentation to optimize production [26].

Design-Build-Test-Learn cycles

A better understanding of cell physiology and technical developments accelerate the design of cell factories and bioprocesses. For instance, when penicillin production programs started in the '60s, strain improvement was based on random mutagenesis followed by analysis and selection of superior strains [27]. With the development of molecular biology, directed genetic modifications became possible. At the same time, detailed information about the organization of the β-lactame genes, responsible for penicillin synthesis, enabled the rational selection of engineering targets. These advances resulted in a 176% increased productivity, corresponding to about five years of strain improvement based on random mutagenesis [27]. However, rational solutions to improve production pathways still require very long development times that limit their application [28]. For instance, the economical production of artemisinin in yeast accounted for over 150 person-years worth of work, and approximately 15 years and 575 person-years were needed to develop and produce 1,3-propanediol by DuPont [28, 29]. Although these products have reached commercialization, only one in 5,000 to 10,000 innovations in industrial biotechnology make the long route from initial finding to market implementation, and efficient approaches for strain and bioprocess development are needed [30].

Design-Build-Test-Learn (DBTL) cycles are a systematic approach to iteratively improve the performance of a biological system (Figure 1.2) [31, 32]. In the *design* phase, researchers plan the genetic modifications required to achieve a desired function, as well as the environmental conditions to be used during testing. In the *build* phase, the designed genetic constructs are

implemented in the organism using synthetic biology tools to insert, delete, or modify genetic material. The engineered strain is then *tested* to evaluate its performance. This could involve assessing the strain's ability to produce a desired product or tolerate specific environmental conditions. In the *learning* phase, the results obtained during testing are analyzed to understand the impact of the genetic modifications or environmental conditions on the performance of the cell factory. The information gathered is then used to inform the next design phase.

In the last decades, technological advances have accelerated all the phases of the DBTL cycle (Figure 1.2) [15]. To name a few, the combination of multi-part DNA assembly techniques [33], the standardization of plasmid structures [34], and the development of genome engineering technologies including CRISPR/Cas [35] have facilitated the strain construction process. In the test phase, the general use of omic techniques has driven systems metabolic engineering, enabling quantitative and qualitative analysis of each regulation layer in a cell [36, 37, 38]. The use of biosensors, able to translate metabolite concentrations to easily measurable signals, allows online monitoring of cell performance [39, 40]. This can be combined with controlled mini-bioreactor systems and high-throughput analytics [41, 42]. The design phase has benefited from the characterization and standardization of genetic parts [43, 44, 45], as well as the creation of databases such as KEGG [46], Brenda [47], or MetaCyc [48], especially for the first iterations of the DBTL cycle [31]. When sufficient information is known about the systems, the use of computational tools such as kinetic, constrained-based, or machine learning models, trained during the learning phase, can guide sequential design phases, effectively linking subsequent cycles.

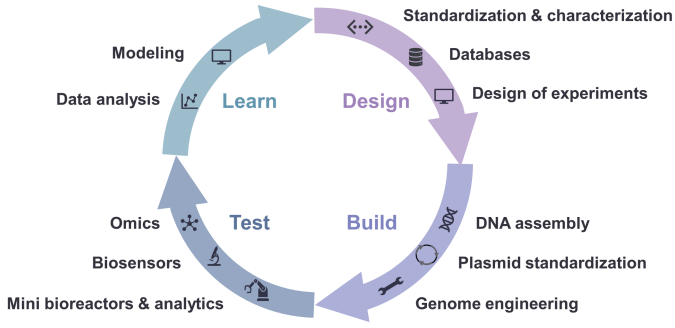


Figure 1.2: Example of technological advances that accelerate Design-Build-Test-Learn cycles.

Biofoundries are highly automated facilities that enable rapid and efficient DBTL cycles (or parts thereof) so cell factories can be built and tested at increasing throughput [49]. However, the acceleration of the design and build phases of the DBTL cycle will only lead to improvements in the performance of cell factories when efficient, meaningful links between the design and learning phases are established [50]. For this, the integration of computational methods, which can be used to improve the understanding of the studied system and/or to guide its optimization, is crucial [51]. The integration of all the phases of the DBTL cycle will ultimately lead to automated systems

that drive experiments, optimize production processes, and facilitate technology transfer and tasks ranging from data gathering to problem-solving [26]. This can lead to a disruptive change that makes the production of new molecules a relatively easy task minimizing the required time for successful strain and bioprocess design [52].

Model-guided Design and Learn

Mathematical models are simplified representations of reality. We model everything, from economics to elections, climate change, or the coronavirus pandemic [53]. Although models are not always right, they are used to model the stock market, the spread of a disease, or to predict tomorrow's weather [53].

Models are constructed based on available data and our understanding of the modeled system. Even when we cannot completely comprehend reality, as is the case with microbial cell factories, models help us learn, organizing our knowledge and unveiling knowledge gaps. Although mathematical modeling is not completely integrated into most metabolic engineering efforts, multiple modeling approaches can be applied to guide the design of cell factories and bioprocesses and to perform data analysis during the learning phase of DBTL cycles [15]. Which model to use (or build) depends on the problem to solve, the experimental factors that can be changed, and the data that can be gathered [54]. In a broad sense, models can be used to enhance the understanding of the studied system or to optimize it. For example, a model can be used to obtain a detailed description of a metabolic pathway and its enzymes or to find conditions that improve the production of a relevant metabolite [54]. However, both objectives are related: a better understanding of a system facilitates its optimization, and the identification of factors with a relevant impact on production can be used to prioritize the aspects of the system that should be further studied.

Knowledge-based models: focus on understanding

Mechanistic models describe the behavior of a system in terms of its biological components and their interactions, boosting interpretability, transparency, and explainability. They provide a rational and systematic framework for integrating existing knowledge and experimental data that allows the validation of model assumptions or the identification of knowledge gaps [31, 54, 55]. In industrial biotechnology, these models are often used to describe production pathways, the metabolism of a cell factory, or its behavior in a reactor context. Depending on the modeling objective, different levels of abstraction might be used. For example, the metabolism of a microorganism might be explained in detail using mass-action kinetics describing the interaction between enzymes and metabolites, it can be simplified using models only based on reaction stoichiometry or further reduced to a single equation [56]. While kinetic models can be used to study pathway dynamics in detail [55], stoichiometry-based models are useful to study cell metabolism [57] and Herbert-Pirt equations are often used in the context of bioprocesses [58].

Here we will focus on kinetic pathway models and genome-scale constraint-based models (Figure 1.3). While both modeling approaches are described by a set of biochemical reactions, kinetic models explicitly describe reaction fluxes as a function of metabolites and enzyme concentrations and enable dynamic simulations and the integration of regulatory information. In turn, constraint-based models solely rely on reaction stoichiometries and assume steady-state conditions. The reader is referred to Saa et al. [55] and Carter et al. [59] for a detailed explanation of these modeling approaches.

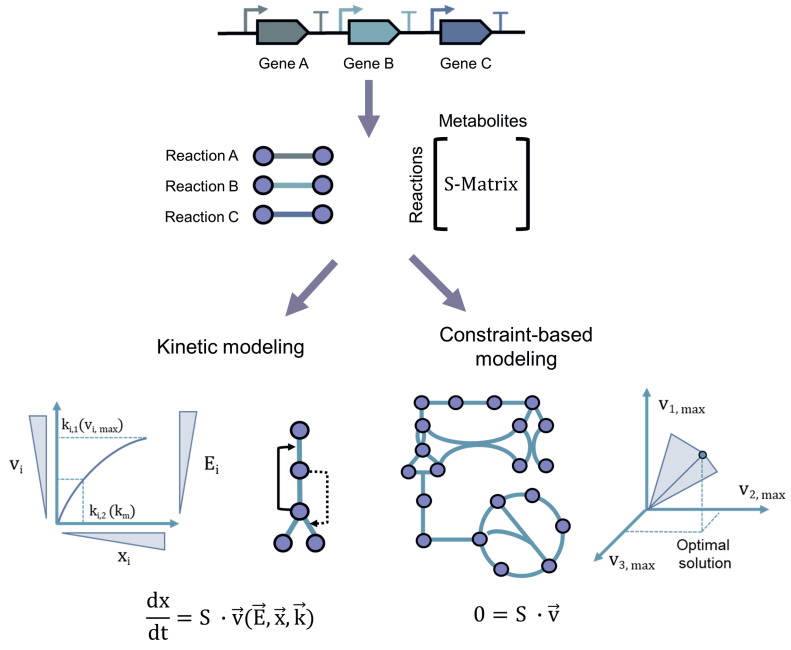


Figure 1.3: Kinetic Modeling and Constraint-based Modeling. Both modeling approaches describe a biochemical reaction network based on the expression of genes, whose proteins perform reactions. The stoichiometric information of the network is stored in the stoichiometry matrix (S-matrix), which can include information on a specific pathway or the complete metabolism. In kinetic modeling, the changes in metabolite concentration (x) over time depend on the S-matrix and the reaction fluxes (v). In turn, reaction fluxes are a function of enzyme concentration (E), metabolite concentrations (x), and kinetic parameters (k) such as maximum fluxes (v_{max}) and Michaelis-Menten constants (k_m). Constraint-based modeling assumes a lack of accumulation of metabolites (steady state), which allows the calculation of reaction fluxes (v) using linear programming by establishing a cellular objective and minimum and maximum flux bounds.

Kinetic modeling

Kinetic models describe a metabolic system through reaction rates based on the integration of stoichiometric, thermodynamic, kinetic, and regulatory information (Figure 1.3) [55]. These models are especially relevant when studying dynamic systems where metabolic regulation, toxicity, or non-linear kinetics play a significant role [54]. The dynamic behavior of a metabolic network is represented by a system of Ordinary Differential Equations (ODEs) that describes the metabolite mass balances:

$$\frac{dx}{dt} = S \cdot v(E, x, k), x(0) = x_0 \tag{1.1}$$

where S and v denote the stoichiometry matrix reconstructed from genomic information and the vector of metabolic reaction fluxes, respectively. Reaction fluxes (v) are a function of enzyme concentrations (E), metabolite concentrations (x), and kinetic parameters (k) (Figure 1.3) [55].

Kinetic models require the identification of appropriate reaction mechanisms which can be explained using reaction rates such as Michaelis Menten kinetics [51, 60]. These rates do not only need to explain the mechanism of the enzyme kinetics but should also be simplified based on the availability of data for model parameterization [31, 60]. The parameterization process requires high amounts of experimental data including enzyme kinetics, enzyme and metabolite concentrations, and thermodynamic information [31, 51]. Although *in vitro* parameter values can serve as an approximation for *in vivo* values, discrepancies between these sources are common [60]. Since the complexity of the parameterization process and the simulation time increases with model size, kinetic reconstructions are often limited to specific pathways, and model reduction approaches have been developed [31, 54].

Parameterized models can be used to simulate pathway dynamics [31]. The effect of tuning enzyme levels or allosteric regulators can be assessed, and reactions operating close to equilibrium can be identified [55]. Alternatively to the use of fully parameterized models, ensemble models can be employed. These models might include different biochemical models describing alternative regulatory mechanisms or a single model structure parameterized by different parameter values [55]. Ensemble models can then be used to study system properties, such as model sensitivity to parameters, while reducing experimental data requirements [61].

Constraint-based modeling

Genome-scale metabolic models (GEMs) are mathematical representations of the complete (known) cell metabolism [57]. Information about the biochemical network of an organism is obtained from its genome following gene-protein-reaction (GPR) rules. Genes codify enzymes that catalyze biochemical reactions with a known stoichiometry that can be retrieved from curated databases such as KEGG [57]. Similarly to kinetic models, the stoichiometry information is stored in the stoichiometry matrix (S) which must be mass and charge balanced [51, 57]. However, GEMs assume a steady state metabolism, overcoming the challenges of scale and data availability characteristics of kinetic models (Figure 1.3). This simplifies Equation 1.1 to:

$$\frac{dx}{dt} = 0 = S \cdot v, \tag{1.2}$$

where the reaction rate (v), usually referred to as flux, is no longer a function of enzyme or metabolite concentrations, nor kinetic parameters, but is still subject to thermodynamic constraints that determine reaction directionalities [51]. Moreover, the ODE problem characteristic of kinetic models is substituted by a set of linear equations that can be solved using optimization techniques such as linear programming (Figure 1.3) [31].

Constraint-based models predict phenotypes in terms of growth and production rates or distributions of metabolic fluxes, often calculated using Flux Balance Analysis (FBA). This technique requires determining an objective function to solve the under-defined system of linear equations given the steady-state assumption, and constraints regarding reaction reversibility and nutrient availability [62]. A widely used objective function is the optimization of the growth rate, represented by the biomass reaction, composed of the essential metabolites needed for growth [62]. However, cellular objectives are found to be condition-dependent and the selection of growth is not always realistic [63]. Additionally, due to the size of GEMs, flux solutions obtained from FBA are often not unique, as multiple flux profiles can achieve the same optimal objective value [54]. As an alternative to FBA, flux sampling aims to explore the entire feasible flux space without imposing an objective, and provides uncertainty margins to the predicted flux distribution [64].

The scope of GEMs has been expanded to include additional thermodynamic constraints [65] or resource allocation strategies such as limitations on membrane surface area or cell volume [66], improving flux predictions. For instance, the GEM with Enzymatic Constraints using Kinetic and Omics (GECKO) framework generates enzyme-constrained models (ecGEM) adding additional constraints linked to the limited enzyme production capacity of the cell and enzyme turnover numbers [67]. Additional layers of cell physiology such as transcription and signal transduction are further included in models such as metabolic and expression models (ME models) [68].

Given their broad scope, GEMs are commonly applied to suggest metabolic engineering targets to enable the overproduction of a metabolite of interest, explore the potential of different metabolic pathways, investigate the impact of integrating new pathways in metabolism, or calculate maximum theoretical yields [54, 59, 69]. However, the steady-state assumption implies that these models are only valid under exponential growth or in continuous cultures with fixed growth rates [51]. Alternatively, dynamic FBA (dFBA) extends the FBA framework by introducing kinetic equations for extracellular metabolites and biomass [70]. This method considers changes in cell growth and metabolism as a response to changes in environmental conditions and assumes steady state for intracellular metabolism, and dynamic changes for the extracellular environment [70].

Data-driven models: focus on optimizing

To build mechanistic models, a deep understanding of cellular processes is needed [50, 52]. Alternatively, data is the main driving force behind data-driven models, which can capture enough of the relevant relationships of the system under study without requiring prior knowledge of its underlying processes [54]. While mechanistic models require assumptions to simplify the relationships between model inputs and outputs, data-driven models can capture the complexity of biological data [52]. These models can be used to study multi-omic datasets, identify inputs with a relevant impact on model output, or guide experimental design [31]. Although these models are often not interpretable, they facilitate the extraction of information from big datasets and point at relevant components or interactions, improving biological knowledge [31].

Considering the numerous factors that affect the performance of a cell factory, the number of experiments required to identify the interplay between these factors is larger than the experimental capabilities for engineering and screening typically found in biofoundries [51]. Therefore, we will focus on the use of statistical design of experiments (DoE) and machine learning (ML) to guide the adjustment of gene expression and/or operational conditions to optimize the production of a target metabolite. These methods link the design and learning phases of the DBTL cycle but differ in the requirements for data generation as well as in the information gained after experimentation (Table 1.1).

For a detailed explanation of DoE and its application to synthetic biology, the reader is referred to Kumar et al., Lawson et al., and Gilman et al. [71, 72, 73]. A general review of ML concepts, algorithms, and their applications beyond bioprocess optimization is presented in Asnicar et al., Volk et al., and Lawson et al. [52, 74, 75].

Table 1.1: Designs of Experiments and Machine Learning.

	Design of experiments		**Machine learning (ML)**	
	Screening	**Optimization**	**Exploration**	**Exploitation**
Library generation	Designed		Designed / Random	
Type of factors	Categorical / Numerical	Numerical	Categorical / Numerical	
# Levels	2	>2	Any	
# Experiments	$n_{factors}^{n_{levels}-k}$		>5% library	
Information gain	Main effects, interactions	Direction of optimum	Explainable ML	

Statistical design of experiments

Design of Experiments (DoE) allows an efficient and structured exploration of a large design space to evaluate the effect of a variety of factors which can take different values (levels) in the system's response [51, 73]. DoE involves the selection of experiments to perform, the use of these experiments to train (linear) models, and the statistical evaluation of the model coefficients.

The simplest DoE design is the full factorial design, which requires performing experiments that include all possible factor and level combinations. Given F factors and L levels, the number of experiments can be calculated:

$$Experiments = \prod_{i=1}^{F} \mathsf{L}_i \tag{1.3}$$

In this design, the effect of a factor, defined as its main effect, is calculated considering the average of all the experiments where the level of the factor is constant, regardless of the other factors. Similarly, interactions between factors are obtained considering experiments where the levels of the studied factors change, regardless of the other factors. When performing a full factorial design, all combinations of factors and levels are tested and the conditions that result in an optimal response are found. Moreover, an analysis of variance (ANOVA) is employed to determine which factors and interactions have a significant effect on the response variable. The degree to which the factors and their interactions affect the response is stored in the model coefficients [71, 72].

The main drawback of full factorial designs is the exponential increase of experiments required to test multiple levels and factors. Instead, fractional factorial designs reduce the number of experiments to perform and allow screening multiple factors and identifying those with the highest impact on the response [71, 72, 73]. Fractional factorial designs test two levels per factor and, depending on their resolution, provide information on main effects and/or interactions. Since multiple experiments are employed to estimate main effects and interactions, these designs reduce the number of experiments to perform maintaining the capacity to estimate some of the model parameters by performing experiments that preserve orthogonality in the desired factors. This ensures that the effect of a factor is not confounded by planned changes in other factors.

Alternatively, Optimal Experimental Designs (OED) are useful to select experiments tailored to any potential experimental constraints, model structure, and estimated parameter values [73, 76]. For example, D-optimal designs allow the design of experiments that improve specific model parameter estimates by minimizing the determinant of the parameter covariance matrix [51, 72]. Notably, these designs are not only applicable to regression models but can also be used for the parameterization of mechanistic models [77, 78].

After screening designs, the obtained model can be used to find the best factor combinations given the experimental space defined by the factor's levels or to guide the expansion of the design space using response surface methods. These methods use a quadratic model as an approximation of the relationship between the factor's levels and the response and require increasing the number of tested levels per factor. Experiments are designed to minimize the variance of the predicted values and can be constructed following, among others, central composite designs, Box-Behnken designs, or I-optimal designs [71]. The generated model is then used to predict optimal factor levels with the aid of contour plots, canonical or ridge analysis [72].

The design of an appropriate DoE strategy for pathway and bioprocess optimization depends on the available knowledge about the system [51]. DoE has been advantageously implemented in the manufacturing and chemical industry [49] and it is often used for bioprocess optimization [79, 80, 81, 82, 83, 84, 85]. This method is increasingly employed during pathway design [32, 86, 87]

and can be used to simultaneously optimize environmental and genetic factors [88, 89].

Machine learning

Machine learning (ML) includes a flexible set of tools for identifying (non-linear) relationships between factors and their effect on the response variable [74]. ML algorithms are classified into supervised and unsupervised methods. In unsupervised learning, ML algorithms aim to find unknown structures in the data without any previous knowledge of potential associations among samples (*e.g.* clustering, dimensionality reduction) [52, 74]. During supervised learning, predicted models are trained based on labeled data, *i.e.* data that contains the value of the response as a function of the modeled features. The common workflow to train these models includes their training with a fraction of the available data (train set) and the validation of their performance based on unseen data (test set) [72, 74]. Validated models can then be used to predict the response given new combinations of features. Depending on the nature of the response variable, classification or regression algorithms are employed [75].

In the context of industrial biotechnology, ML has been used for gene annotation and pathway design, pathway building, performance testing, and production scale-up [90, 91, 92, 93, 94]. Pathway optimization usually involves tuning gene expression through the modification of promoters and RBS sequences and can be aided by supervised ML methods [52]. Examples of the use of ML for pathway optimization range from the use of simple linear regression [95] or random forest [96] to artificial neural networks [97]. Besides, algorithms that autonomously suggest a recommended subset of variables to test experimentally based on training data such as BioAutomata [98], ART [99], or METIS [100] are available. These methods require a balance between exploration and exploitation. Exploration is usually prioritized in initial DBTL cycles where experiments that include a diverse range of conditions are suggested to gain a better understanding of the system. Subsequent cycles focus on exploitation so the proposed experiments aim at finding the best-producing strains [31]. While exploration is similar to screening fractional factorial designs, exploitation is equivalent to response surface designs within the DoE methodology.

ML methods are agnostic to the nature of the factors and can easily accommodate factors with different numbers and types of levels. However, they require large, high-quality datasets to avoid bias in the predictions [96]. Although ML algorithms are often used to analyze data obtained by randomly generated libraries, they can also be trained with data obtained using DoE designs, which ensures unbiased sampling of the solution space [101]. Compared to simpler linear models, the ability of ML to accommodate complex datasets results in less accessible information about the studied system. Yet, explainable machine learning techniques are available [102]. For instance, opaque models can be locally approximated by simpler, explainable models or the relevance of the model features can be estimated to identify factors with the biggest impact on the response [102, 103].

Aim and thesis outline

During this thesis I aimed to deploy different modeling strategies to guide and accelerate the design and learning phases of DBTL cycles for strain and bioprocesses development. These modeling approaches were envisioned to facilitate the optimization of strains and bioprocesses, and differ in their requirements regarding *a priori* knowledge about the system (design) and the new information generated after experimentation (learning). Hence, in **Chapters 2 to 5** I focus on the use of knowledge-based models, **Chapter 6** is based on the analysis of omics data, and in **Chapters 7 to 9** I delve into the use of data-driven methods (Figure 1.4).

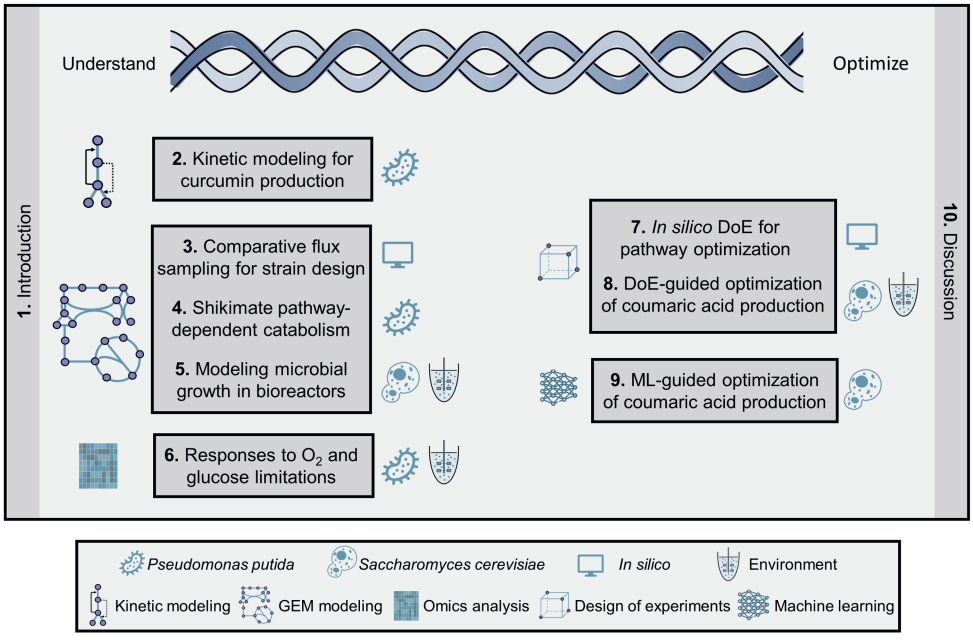


Figure 1.4: Thesis overview. The thesis is structured based on the applied methods and the relative weight given to mechanistic understanding and optimization of the studied system.

During this work, *in silico* studies (**Chapters 3, 7**) were accompanied by experimental work using *Saccharomyces cerevisiae* (**Chapters 5, 8, 9**) and *Pseudomonas putida* (**Chapters 2, 4, 6**). While *S. cerevisiae* is a widely used eukaryotic model organism for the generation of bio-products [104], *P. putida* is a Gram-negative bacteria gaining recognition as a versatile chassis for industrial biotechnology [105]. The use of these hosts highlights the easy generalization of the methods used, emphasizing the potential broad applicability of computational approaches for the acceleration of DBTL cycles.

In **Chapters 2 to 4** I used mechanistic models for metabolic engineering. **Chapter 2** delves into the development of a kinetic model of the curcuminoid production pathway expressed in *P. putida*. This pathway is characterized by the presence of promiscuous enzymes and requires a balanced expression of its genes to improve production. Despite the complexity of the parameterization process, I explored the potential of ensemble model simulations to guide the design of improved strains and identify knowledge gaps in the studied pathway. In **Chapter 3** I present the Comparative Flux Sampling Analysis (CFSA) tool for the identification of down-regulation and over-expression targets for metabolic engineering. This tool is based on the use of GEMs to simulate growth and production scenarios and the identification of reactions with altered fluxes. The use of flux sampling ensures a complete exploration of the solution space, increasing the robustness of the predictions. This tool was applied to improve the production of lipids by *Cutaneotrichosporon oleaginosus* and naringenin production by *S. cerevisiae*. In both cases, experimentally validated targets as well as new engineering strategies were identified. While CFSA is designed to provide growth-uncoupled production strategies, **Chapter 4** focuses on the development of a new-to-nature shikimate-dependent catabolism in *P. putida* for growth-coupled production of shikimate-derived products using a model-guided approach. I employed GEM simulations to design a pyruvate auxotroph strain that relies on pyruvate produced in the shikimate pathway for growth. This strain was constructed and subject to adaptive laboratory evolution coupled with the use of a biosensor to identify evolved strains with high shikimate fluxes. The evolved strain was further engineered for the production of 4-hydroxybenzoate following a model-driven approach that resulted in 89% of the maximum theoretical yield in minimal media.

While in the previous chapters pathways and metabolism were considered in isolation, in **Chapters 5 and 6** I acknowledge the interplay between bioprocesses and cell metabolism. In **Chapter 5** I use dynamic FBA and an enzyme-constrained GEM to simulate *S. cerevisiae* metabolism in different bioreactor contexts. The combination of these methods allowed the prediction of critical dilution rates, ethanol production and consumption, the preferred consumption order of different carbon sources, and the performance of a lactate-producing strain. Additionally, flux sampling was used to study metabolic changes at different growth rates. Compared to the data availability regarding *S. cerevisiae* growth, little was known about the response of *P. putida* to low oxygen concentrations. Therefore, in **Chapter 6** I performed chemostat cultivations of this bacteria at decreasing oxygen availability and analyzed its response at the physiological, transcriptomic, and proteomic levels. We found that glucose-limited cells grown at low growth rates (0.1 h^{-1}) produce pyoverdine at the expense of lower biomass yields. Besides, *P. putida* cells can endure up to 8 days in oxygen-limited growth which results in the up-regulation of genes related to respiration and minimal changes at the proteome level.

In **Chapters 7 to 9** I move from the use of mechanistic models and omic data to data-driven models that prioritize production optimization over mechanistic understanding. In **Chapter 7** I used the kinetic model developed in **Chapter 2** to study the efficiency of different DoE models for pathway optimization. The use of the kinetic model enabled the simulation of a full factorial library that

was used to estimate the effect of factors as well as identify the enzyme concentrations resulting in optimized production. Then, the information obtained by different DoE designs was compared to the full factorial data. Designs providing the required minimum information for optimization while reducing the experimental load were identified. The gained knowledge was applied in **Chapter 8** where I performed a DoE-guided optimization of p-coumaric acid production in *S. cerevisiae*. This chapter highlights the importance of simultaneous process, media, and genetic optimization during the cell factory design process. We emphasized the role of DoE during sequential experimentation and its ability to point at relevant factors for production while the experimental effort is minimized. As an alternative to DoE, **Chapter 9** also delves into p-coumaric acid production in *S. cerevisiae* but focuses on the use of ML for pathway optimization. This approach allowed a flexible design including factors with different number of levels based on prior knowledge that went beyond gene expression tuning to additionally testing the effect of different coding sequences. A library of strains was obtained by one-pot transformation, and a random screening before sequencing approach enabled the generation of high-quality data for model training. I employed ML models to predict the best strain in the original design space and the relevance of model features to guide the expansion of this space resulting in a 68% increased production.

Finally, in **Chapter 10**, I reflect on the lessons learned during this journey. I delve into the successes and limitations of the performed studies and provide a set of recommendations to benefit from different modeling strategies at each step of bioprocess development. I wonder about the complementary roles of understanding and optimization is biotechnology. In addition, I reflect on the potential and challenges of biotechnology when automation and computational modeling are combined into autonomous DBTL cycles. Finally, I leave the reader with some thoughts to reflect on the ability of biotechnology to fulfill its promises in the context of the circular economy.

CHAPTER 2

Model-guided metabolic engineering of curcuminoid production in *Pseudomonas putida*

Sara Moreno Paz*, María Martín Pascual*, Rik P. van Rosmalen*, Julia Dorigo, Francesca Demaria, Richard van Kranemburg, Vitor A. P. Martins dos Santos#, María Suárez Diez#

*Contributed equally, #Jointly supervised this work

This chapter is ready for submission

Abstract

Production of value-added, plant-derived compounds in microbes increasingly attracts commercial interest in food and pharmaceutical industries. However, plant metabolic pathways are complex, require a robust balance of enzymes, cofactors, ATP, and other metabolites, and often result in low production when expressed in bacteria. This is exemplified by the biosynthesis of curcuminoids from the *Curcuma longa* plant. Here, we combine dynamic pathway modeling, systematic testing of isoenzymes, and the optimization of gene expression levels and substrate concentrations for the biosynthesis of curcuminoids in *Pseudomonas putida*, leading to unprecedented conversion rates of caffeic acid and tyrosine to curcumin. The development of kinetic ensemble models guided the design of production strains, emphasizing the necessity of high relative expression of *c3h*, *curs2*, and *dcs* and the low relative expression of *tal*, *comt*, *ccoaomt*, and *4cl4*. This optimization resulted in a strain that achieved $10.8 \pm 1.8\%$ of the maximum theoretical yield of curcumin from tyrosine. This represents a 4.1-fold increase in production efficiency and the highest yield reported to date, demonstrating the potential of *P. putida* as a platform for curcuminoid production. Our findings highlight the effectiveness of this strategy not only in the advances in the production of curcuminoids but also in setting a framework for the biosynthesis of other complex compounds.

Introduction

Curcuminoids are polyphenolic compounds naturally found in the rhizome of the *Curcuma longa* (turmeric) plant. They account for 1 to 6% (w/w) of the turmeric rhizome, with curcumin being the most abundant of the total curcuminoids (60-10% w/w), followed by demethoxycurcumin (20-27% w/w) and bisdemethoxycurcumin (10-15% w/w) [106]. Due to their characteristic strong yellow color, curcuminoids are widely used as coloring agents in the food industry and have been authorized as food additives by the European Union (code E100) and the FDA (label 73.615) [107, 108]. Besides their use in the food industry, the pharmaceutical and cosmetic industries account for 50% of the curcuminoids global market [109]. Multiple biological activities, predominantly anti-carcinogenic, antioxidant, and anti-inflammatory, have been attributed to curcumin and its derivatives but further research is still needed to ascertain their full therapeutic potential [106, 110].

The curcumin global market size was over USD 58 million in 2020, and it is predicted to grow at a compound annual growth rate (CAGR) of 16.1 % by 2028 [109]. Meeting such global demands with plant-based curcumin production is challenging due to the seasonal dependent growth of the turmeric plant, hindered by the expected rise of temperatures [111]. Furthermore, traditional methods for the extraction of curcuminoids from the rhizome of the plant are lengthy, laborious, and require the use of organic solvents, high pressure, and elevated temperatures. Although alternative methods hold promise for green extractions, they are inefficient, only able to extract up to 6% of the curcuminoids present in the rhizome [112]. Chemical synthesis of curcuminoids also has severe downsides, including the reliance on fossil fuel-derived solvents, toxic reagents, and expensive starting compounds [113]. Microbial production of curcuminoids provides a valuable alternative to these processes, as it has the potential to provide a greater level of control, consistency, and efficiency, whilst operating at milder conditions and relying on biomass-derived feedstocks such as glucose, tyrosine, ferulic, p-coumaric and caffeic acids. This allows, in principle, for efficient, biobased, and standardized large-scale production of curcuminoids with the potential to meet the increasing global demands.

The curcuminoid pathway has been expressed in various microorganisms, with a strong focus on *Escherichia coli*, where the maximum reported yields have been obtained (Sup. Table 2.1). The conversion of ferulic acid to curcumin requires only three catalytic reactions, and 100% yields have been achieved [114]. Although the same enzymatic steps are required for production from p-coumaric acid, for this substrate, only 59% of the maximum theoretical yield has been reported [115]. When caffeic acid or tyrosine are used as substrates, the highest curcumin yields decrease to 2.12 and 1.27 % of the maximum theoretical yield, respectively [115, 116]. These lower yields can be caused by the increased complexity of the pathway or the higher toxicity of these substrates and results in not yet commercially viable bio-based production of curcuminoids.

The curcuminoid biosynthetic pathway has been reconstructed based on curcuminoid production in *C. longa* and *Oryza sativa* and expanded to include enzymatic reactions derived from various yeast, bacterial, and plant species (Figure 2.1) [115, 117]. Curcuminoids can be formed from two tyrosine molecules in several steps. First, tyrosine is deaminated to form p-coumaric

acid by tyrosine ammonia lyase (TAL). Next, p-coumaric acid can be converted to caffeic acid by coumarate-3-hydroxylase (C3H) or to coumaroyl-CoA by feruloyl/coumaroyl-CoA synthase (FCS) or 4-coumarate-CoA ligase (4CL). Subsequently, caffeic acid can be directly converted into ferulic acid by caffeic acid O-methyl transferase (COMT). Alternatively, it can be ligated to coenzyme-A (CoA) by FCS or 4CL, creating caffeoyl-CoA, which later can be converted to feruloyl-CoA by caffeoyl-CoA O-methyl transferase (CCOAOMT). When ferulic acid is formed, it can also be bound to CoA by FCS or 4CL, creating feruloyl-CoA. Hence, the three hydroxycinnamic acids (ferulic, p-coumaric and caffeic acids) can be ligated to CoA by the action of either FCS or 4CL. Once the CoA-esters coumaroyl- and feruloyl-CoA are formed, malonyl-CoA is added to these compounds in a condensation reaction catalyzed by diketide-CoA synthase (DCS). Finally, the diketide-CoA esters react with a single molecule of a CoA ester, coumaroyl-CoA or feruloyl-CoA, to form a curcuminoid. This last reaction step is catalyzed by curcumin synthase (CURS). When feruloyl-diketide-CoA reacts with a molecule of feruloyl-CoA, curcumin is formed. When coumaroyl-diketide-CoA reacts with coumaroyl-CoA, bisdemethoxycurcumin is formed, and when either diketide-CoA reacts with one of the other CoA esters (*i.e.* feruloyl-diketide-CoA with coumaroyl-CoA or vice versa), the asymmetric demethoxycurcumin is formed. There are three subtypes of CURS with different substrate specificity. CURS1 and CURS2 prefer feruloyl-CoA as a starting substrate, and CURS3 does not show any preference between these metabolites [117]. Furthermore, the last two reactions, catalyzed by DCS and CURS, can also be performed by a single enzyme, curcuminoid synthase (CUS) that, although has a preference for coumaroyl-CoA and the production of bisdemethoxycurcumin, is also able to produce curcumin in trace amounts [115].

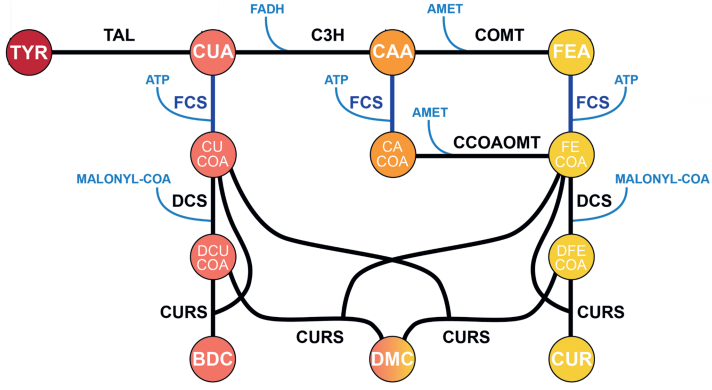


Figure 2.1: Overview of the curcuminoid production pathway showing the potential substrates and products. Metabolite abbreviations: TYR, tyrosine; CUA, p-coumaric acid; CAA, caffeic acid; FEA, ferulic acid; CUCOA, coumaroyl-CoA; CACOA, caffeoyl-CoA; FECOA, feruloy-CoA; DCUCOA, diketide coumaroyl-CoA; DFECOA, diketide feruloyl-CoA; BDC, bisdemethoxycurcumin; DMC, demethoxycurcumin; CUR, curcumin; AMET, S-adenosyl-l-methionine. Enzyme abbreviations: TAL, tyrosine ammonia lyase; C3H, coumarate-3-hydroxylase; COMT, caffeic acid O-methyl transferase; FCS, feruloyl/ coumaroyl-CoA synthase; CCOAOMT, caffeoyl-CoA O-methyl transferase; DCS, diketide-CoA synthase; CURS, curcumin synthase.

The presence of promiscuous enzymes in the curcuminoid pathway complicates its optimization, as alterations in gene expression can lead to unforeseen effects on reaction rates. To address this challenge, dynamic pathway models, which are constructed using a system of ordinary differential equations (ODEs), can be employed. These models account for metabolite concentrations and enzyme fluxes over time. They rely on mathematical representations of enzyme kinetics, such as Michaelis-Menten equations, and necessitate the determination of kinetic parameters [118, 119]. They serve as a valuable tool for comprehending pathway fluxes, assessing (non-measurable) intermediate concentrations, and pinpointing potential limiting steps. Furthermore, they can be used to evaluate how alterations in enzyme concentrations impact production, thereby aiding in the design of optimized pathways [118, 119].

To enhance curcuminoid production, the choice of a microorganism able to endure toxic pathway metabolites is essential. *Pseudomonas putida* KT2440 has gained recognition as a promising platform for the production of various biological products. It offers a robust metabolism as well as a high tolerance to a variety of substances, particularly aromatic compounds, making it a suitable candidate for tolerating the toxicity of the various substrates and intermediates in the curcuminoid pathway [105, 120, 121, 122]. Natively, *P. putida* is able to catalyze the conversion of hydroxycinnamic acids into CoA ester molecules due to the presence of *fcs* (PP_3356). However, it can degrade the phenylpropanoid-CoA metabolites using ECH (PP_3358), which interferes with the production of curcuminoids. Although production of bisdemethoxycurcumin from p-coumaric acid has been achieved in *P. putida* Δech, only a 0.2% yield has been obtained, and the potential of this host for curcuminoids production remains largely unexplored [123].

Here, we demonstrate the successful production of curcuminoids from ferulic acid, p-coumaric acid, caffeic acid, and tyrosine facilitated by plasmid-based expression of heterologous genes in *P. putida* Δech. A curcumin yield of $63.6 \pm 3.0\%$ of the maximum theoretical yield was reached from ferulic acid, $24.0 \pm 8.7\%$ from p-coumaric acid, $48.5 \pm 9.1\%$ from caffeic acid, and $2.7 \pm 0.2\%$ from tyrosine. Experimental data were used to create ensemble dynamic models of the curcuminoid pathway that connect enzyme concentrations, enzyme kinetics, thermodynamics, and metabolite concentrations. These models served as a guide in designing new strains with varying expression levels of the pathway genes, resulting in the optimization of curcumin production from tyrosine up to a $10.8 \pm 1.8\%$ of the maximum theoretical yield, representing a 4.1-fold increase in production. This study highlights the potential of dynamic pathway models in comprehending intricate production pathways, providing a foundation for generating hypotheses that guide the optimization process. Moreover, the curcumin yields achieved from caffeic acid and tyrosine stand as the highest reported to date. Additionally, curcumin was successfully produced from glucose, showcasing the potential of *P. putida* as a promising cell factory for the production of curcuminoids.

Materials and methods

Media, Bacterial strains and chemicals

Lysogeny-Broth (LB) (10 g/l tryptone, 10 g/l NaCl and 5 g/l yeast extract) and M9 minimal media (1.63 g/l NaH_2PO_4, 3.88 g/l K_2HPO_4, 2 g/l $(NH_4)_2SO_4$, 10 mg/l EDTA, 100 mg/l $MgCl_2 \cdot 6H_2O$, 2 mg/l $ZnSO_4 \cdot 7H_2O$, 1 mg/l $CaCl_2 \cdot 2H_2O$, 5 mg/l $FeSO_4 \cdot 7H_2O$, 0.2 mg/l $Na_2MoO_4 \cdot 2H_2O$, 0.2 mg/l $CuSO_4 \cdot 5H_2O$, 0.4 mg/l $CoCl_2 \cdot 6H_2O$, and 1 mg/l $MnCl_2 \cdot 2H_2O$) were used to grow the bacterial strains. When necessary, antibiotics were added to the media: kanamycin (50 µg/ml), gentamycin (15 µg/ml), streptomycin (100 µg/ml), or chloramphenicol (34 µg/ml). *E. coli* λpir competent cells were prepared and used for plasmid propagation [124]. *E. coli* cells were grown in LB media at 37°C in a shaking incubator at 200 rpm. *P. putida* cells were grown in LB or M9 media at 30°C in a shaking incubator at 200 rpm. The *P. putida* Δ*ech* strain was generated from *P. putida* KT2440 using the pGNW and pQURE plasmids [125]. *P. putida* Δ*ech* was made electrocompetent after several washing steps with 300 mM sucrose. A single exponential decay pulse was applied using the GenePulser Xcell™ (Bio-Rad) set at 2500, 200, and 25 V.

A list of all strains, plasmids, and primers used in this study can be found in Sup. Data 1. Analytical standards of l-tyrosine, ferulic acid, p-coumaric acid, caffeic acid, curcumin, demethoxycurcumin, and bisdemethoxycurcumin, as well as ethyl acetate and DMSO were purchased from Sigma-Aldrich. Acetonitrile and hydrochloric acid were purchased from Acros Organics.

Toxicity essays

Toxicity assays were performed with the four precursors (ferulic, p-coumaric, caffeic acids, and tyrosine) on *P. putida* Δ*ech* in a concentration range of 0-0.25-0.5-1-2-4-8 mM. *P. putida* Δ*ech* was grown overnight at 30°C in a 50 ml falcon tube containing 10 ml of LB. The next day, the culture was centrifuged for 10 min at 4700 g. The supernatant was discarded, and the cell pellet was resuspended in 10 ml of M9 media containing 70 mM of glucose. The OD600 of the cultures was measured in a spectrophotometer (IMPLEN Westburg). A 96-well plate was prepared by adding 200 µl of the sample per well with a starting OD600 of 0.2. The OD600 readings were monitored in a microplate reader (BioTek Synergy/neo2 or BioTek SynergyMx) every 5 minutes for 48 hours at 30°C with a continuous shake.

Plasmid construction

Gene sequences of *tal*, *c3h*, *ccoaomt*, *dcs*, *curs1*, and *comt* were obtained from Rodrigues et al. [116]. Gene sequences of *curs2* and *curs3* were obtained from NCBI (accession numbers AB506762 and AB5067630). All the gene sequences were codon optimized for *P. putida* KT2440 using the Jcat codon optimization tool (www.jcat.de), and genes were ordered from Twist Bioscience. The genes for *curs1* and *dcs* codon-optimized for *C. glutamicum* were kindly provided by Dr. Katarina Cankar (Wageningen Plant Research). Gene sequences can be found in Sup. Data 1.

The genes were first ligated into the pSB1C3 repository plasmid via Golden Gate cloning using BsaI restriction enzyme. The BsaI overhangs were already present on the genes codon optimized for *P. putida*. The *curs* and *dcs* genes codon-optimized for *C. glutamicum* were PCR amplified with M526-M27 and M528-M529 pair of primers, respectively, using NEB Q5® High-Fidelity DNA polymerase according to the manufacturer's protocol. Next, PCR fragments were loaded onto 1% agarose gel and purified using the Nucleospin™ Gel and PCR Clean-up kit from Fischer Scientific™. Plasmids were built using the SevaBrick Assembly method [126] and transformed into *E. coli* λpir competent cells by heat-shock. Colonies were screened for the correct assembly of the plasmid by colony PCR using Phire Green Hot Start II polymerase. Colonies with the right size band were grown overnight in 10 ml of LB supplemented with the appropriate antibiotic. Glycerol stocks were prepared for long-term storage. Plasmids were purified from the overnight liquid cultures using the GeneJET Plasmid Miniprep kit from Thermo Scientific™. Plasmid sequences were confirmed by Sanger sequencing from Macrogen (MACROGEN Inc. DNA Sequencing Service, The Netherlands).

Plasmids carrying different combinations of the pathway genes were created ligating genes into pSEVAb22, pSEVAb83, or pSEVAb25 plasmid backbones carrying a p100 promoter [126]. Combinations of genes and plasmid backbones are listed in Sup. Data 1.

Curcuminoids production experiments

Strains were grown in 10 ml LB medium containing the appropriate antibiotic, at 30°C and 200 rpm, overnight. The next day, the overnight liquid cultures were centrifuged for 10 min at 4700 g. The supernatant was discarded and the cell pellet was washed in 1 ml of M9 media containing 70 mM glucose to eliminate LB traces. Cells were resuspended to a starting OD600 of 0.3 in a total volume of 25 ml of fresh minimal M9 media supplemented with 70 mM of glucose, the appropriate antibiotics, and the precursor (ferulic -, p-coumaric -, caffeic acid, or tyrosine) at indicated concentrations. The 250 ml-Erlenmeyer flasks, containing the cells were grown aerobically at 30°C and 200 rpm for 72 h. Samples were taken for each flask at indicated time points and phenylpropanoids and cucuminoids were extracted. Biological triplicates were included.

Extraction of phenylpropanoids and curcuminoids

To measure curcuminoids and precursors, a sample of 500 µl or 1 ml was taken from each Erlenmeyer, and transferred to a clean 2 ml Eppendorf tube. Then, 1 µl of 6M HCl was added to each tube, and the tubes were vortexed to break the cells. To extract the curcuminoids, an equal amount of ethyl acetate was added to the Eppendorf. The tubes were then incubated at 55°C for 10 min at 800 rpm. Then, the tubes were centrifuged in a microcentrifuge at 20238 g for 2 min. The top layer was transferred to a new tube, making sure none of the cell pellet was taken in the process. This extraction method was repeated until there was no yellow color visible in the cell pellet. The ethyl acetate was then evaporated in a rotary evaporator (Concentrator plus, Eppendorf) at 60°C. Finally, the remaining dry sample was dissolved in 500 µl of DMSO.

Quantification of phenylpropanoids and curcuminoids

High-performance liquid chromatography (HPLC) was used to quantify the substrates (ferulic, p-coumaric, caffeic acid, and tyrosine) and curcuminoids (demethoxycurcumin, bisdemethoxycurcumin, and curcumin). HPLC was performed on a Shimadzu LC2030C machine equipped with a Poroshell 120EC-C18 column (250 x 4.6 mm, Agilent) and a UV/vis detector. Mobile phase was used at a rate of 1 ml/min and was composed of Milli-Q water (A), 100 mM formic acid (B), and acetonitrile (C) at varying proportions: 77:10:13 (v/v/v) in the first 10 min, 23:10:67 (v/v/v) in the next 9 min, and 77:10:13 (v/v/v) in the last six minutes. Curcuminoids and hydroxycinnamic acids were detected at a wavelength of 420 nm and 280 nm, respectively.

Yields were calculated by dividing the maximum measured curcuminoid concentration by the consumed substrate. Yields are expressed as a percentage of the maximum curcuminoid yield calculated using the number of carbons in the substrates (tyrosine, p-coumaric, caffeic, and ferulic acids) and products (bisdemethoxycrucmin, demethoxycurcumin, and curcumin). The maximum curcuminoid yield is 0.5 mol curcuminoid/ mol substrate.

Construction of kinetic models of the curcuminoid pathway

Three models, mFeCua2, mCaa2, and mTyr2, corresponding to strains FeCua2, Caa2, and Tyr2, were developed based on the genes expressed in each strain (Table 2.1). For all reactions but C3H, the generalized reversible Michaelis-Menten kinetics expression rate law was applied. For the C3H reaction, a mass-action rate law was used due to its high equilibrium constant [127]. The rate laws were formulated using the Wegscheider-compliant parametrization, which models the reaction equilibrium constant by considering the contributions of the concentration and chemical potential of each reactant. This approach helps to avoid hidden dependencies between the equilibrium constants of multiple reactions, which could lead to thermodynamically infeasible parameterizations [127]. Enzymes catalyzing multiple reactions (FCS, DCS, and CURS) had additional substrate competition terms incorporated into their rate laws [127]. To simplify the model, the concentrations of all co-factors were fixed, but their chemical potential contributions to relevant rate laws were considered. Additionally, enzyme concentration and rate parameters were merged when possible to mitigate issues related to identifiability. Growth was integrated into the model through a logistic equation, which was used to scale enzyme and metabolite concentrations. The most complex model (mTyr2) encompassed 14 reactions and 13 ODEs to represent the dynamic concentrations of 12 metabolites and biomass. This model involved a total of 89 parameters, with 37 representing Michaelis-Menten constants (k^M), 16 representing enzyme concentrations, rates, or a combination of both (u, k^V, and u^V, respectively), 23 representing standard thermodynamic potentials (μ), 9 representing fixed concentrations of co-factors (c), 2 being part of the logistic equation representing growth, and 2 fixed parameters (R and T). The generated models and detailed information regarding model construction are available in Gitlab and Sup. Methods, respectively.

Table 2.1: Summary of constructed models.

Model name	Included enzymes
mFeCua2	FCS, DCS, CURS2
mCaa2	COMT, CCOAOMT, FCS, DCS, CURS2
mTyr2	TYR, C3H, COMT, CCOAOMT, FCS, DCS, CURS2

Parameterization of the kinetic models

To establish initial parameter value ranges, various literature sources were combined. When available, co-factor concentration data was sourced from studies on *P. putida* KT2440 [128]. Otherwise, information was gathered from *Pseudomonas taiwanensis* VLB120 [129] and *E. coli* K12 [130], with preference given to the former. Estimates for chemical potentials were derived from changes in the Gibbs free energy of the reactions, computed using the eQuilibrator API under standard physiological conditions (pH = 7.5, ionic strength = 0.25 M, and pMg = 3) [131]. Enzyme kinetic parameters for DCS and CURS were obtained from Katsuyama et al. [117, 132]. In instances of missing data, broad estimates were used. All parameter data was integrated through parameter balancing, resulting in a final multivariate normal distribution for the parameters [133]. The covariance between the estimated chemical potentials from eQuilibrator was included in the parameter balancing output, preserving the shared uncertainty in the estimates [134].

For parameter estimation of the mFeCua2, mCaa2, and mTyr2 models, experimental data included measurements of OD600 and concentrations of substrates and curcuminoids from the FeCua2 strain grown in 2 mM ferulic acid or 1 mM p-coumaric acid, the Caa2 strain grown in 1 mM caffeic acid, and the Tyr2 strain grown in 1 mM tyrosine. Parameters were simultaneously estimated for all models using Python (v3.9.12) and the Python libraries PESTO (v0.2.12) [135] and AMICI (v0.11.21) [136]. Initial parameter estimates were sampled from the balanced distribution and log-transformed where appropriate (*i.e.* for parameters representing kinetic constants, enzyme concentrations, or co-factor concentrations). We employed multi-start local optimization (L-BFGS-B, 1000 starts, 100 maximum iterations), with a least-squares objective function that minimized the sum of squares of the residuals between the experimental data and the model predictions. Ensemble models were created for each strain (enFeCua2, enCaa2, and enTyr2), incorporating the top ten parameterized models. The goodness of fit for each observable (tyrosine, caffeic acid, p-coumaric acid, ferulic acid, bisdemethoxycurcumin, demethoxycurcumin, curcumin concentrations, and OD600) in all experiments was evaluated using the mean square error normalized by the mean of the measurements. Simulations were performed using the AMICI library [136] and the CVODES ODE solver [135]. Scripts used for parameterization and simulation are available in Gitlab.

Model-based optimization of curcumin production from tyrosine

Limiting reactions were identified following two approaches. First, ensemble models were used to simulate the conducted experiments and, for each reaction, the maximum predicted flux was calculated. Reactions exhibiting lower flux values were flagged as potential bottlenecks, and the corresponding enzymes were selected as targets for over-expression. Furthermore, the enTyr2 ensemble was used to evaluate the influence of varying enzyme concentrations on curcumin production from tyrosine. The model was simulated under different enzyme concentration scenarios (ranging from 0 to 10 times the predicted concentration, see Gitlab). Based on their impact on curcumin production, enzymes were considered as candidates for either over-expression or down-regulation. Expression of genes in plasmids with distinct copy numbers allowed the experimental implementation of model suggestions. pSEVAb22, pSEVAb83 and pSEVAb25 plasmids were used as low-copy-number, medium-copy-number and high-copy-number plasmids, respectively [126].

Results

Curcuminoid substrate toxicity in *P. putida* Δ*ech*

To establish *P. putida* KT2440 as an efficient platform to produce curcuminoids, degradation of coumaroyl-CoA, caffeoyl-CoA, and feruloyl-CoA by ECH should be avoided. Therefore, we deleted the *ech* gene and evaluated the toxicity of the four precursors (ferulic, p-coumaric, and caffeic acids, as well as tyrosine) on *P. putida* Δ*ech*. In this way, the highest concentration of each precursor that does not inhibit growth was identified. The growth of *P. putida* Δ*ech* was hindered when ferulic and p-coumaric acids were present in the media at concentrations of 4 mM. Similarly, tyrosine at 8 mM and caffeic acid at 0.25 mM completely inhibited growth (Sup. Figure 2.1).

Production of curcuminoids from ferulic and p-coumaric acid

Production of curcuminoids from ferulic and p-coumaric acids requires the expression of two heterologous genes, *dcs* and *curs*, and the native expression of *fcs* (Figure 2.2). Therefore, curcuminoid production from these substrates was considered as the first step towards the expression of the full curcuminoid pathway in *P. putida* Δ*ech*.

Five strains expressing *dcs* and *curs* in a pSEVAb22 plasmid were constructed: FeCua1, FeCua2, FeCua3, FeCuaCg, and FeCuaCgCg (Table 2.2). These strains differed on the *curs* isoenzyme used (*curs1*, *curs2*, or *curs3*) and the codon optimization of *curs* and *dcs* (either for *P. putida* or *C. glutamicum* (Cg)). These strains were grown in M9 supplemented with 1 mM and 2 mM of ferulic acid (Table 2.2). A starting 2 mM concentration of ferulic acid resulted in the highest titers and curcumin production by FeCua2 (0.34 $\pm$ 0.05 mM), followed by FeCuaCg (0.18 $\pm$ 0.09 mM), FeCua1 (0.17 $\pm$ 0.01 mM) and FeCua3 (0.06 $\pm$ 0.02 mM) (Figure 2.2D, Sup. Figure 2.2). FeCuaCgCg grew slower and produced little amount of curcumin compared with the other strains (0.003 $\pm$0.00 mM) (Sup. Figure 2.2). Although FeCua3 was the only strain that accumulated detectable levels

of ferulic acid, the total conversion of ferulic acid to curcumin was always below 100%, lower than previously reported yields in *E. coli* [114]. Moreover, increasing the initial OD600 significantly influenced curcumin production. FeCua2 was inoculated at three different OD600, 0.3, 0.6, and 0.9 with 2 mM of ferulic acid. The highest titer, 0.64 ± 0.03 mM (63.6% of the maximum theoretical yield), was achieved when ferulic acid was added at an OD600 0.9, representing a 1.8 fold-increase compared to the yield obtained with an initial OD600 of 0.3 (Table 2.2, Sup. Figure 2.2C,D).

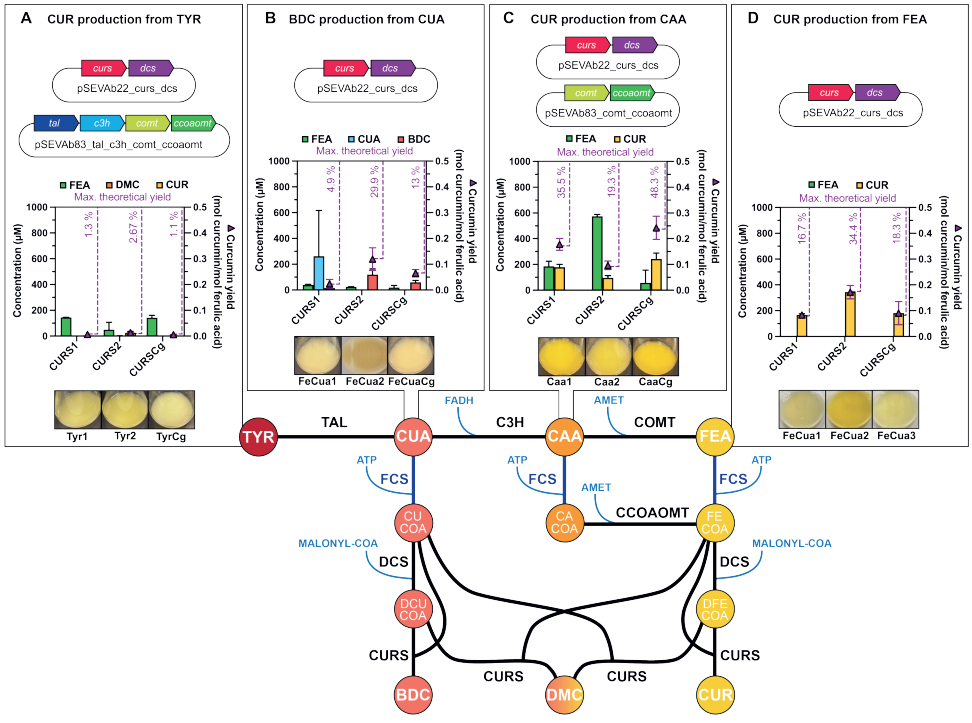


Figure 2.2: **A.** Curcumin (CUR) production from tyrosine (TYR), the performance of the three best strains is shown (see Sup. Figure 2.5 for additional information). **B**. Bisdemethoxycurcumin (BDC) production from p-coumaric acid (CUA), the performance of the three best strains is shown (see Sup. Figure 2.3 for additional information). **C.** Curcumin production from caffeic acid (CAA), the performance of the three best strains is shown (see Sup. Figure 2.4 for additional information). **D.** Curcumin production from ferulic acid (FEA), the performance of the three best strains is shown (see Sup. Figure 2.2 for additional information). See Figure 2.1 for abbreviations.

FeCua1, FeCua2, and FeCuaCg were also grown in M9 media supplemented with 1 mM of p-coumaric acid (Figure 2.2 B, Table 2.2, Sup. Figure 2.3). Similar to the experiments with ferulic acid, FeCua2 produced the highest concentration of bisdemethoxycurcumin (0.12 ± 0.04 mM). Notably, traces of ferulic acid were detected in the samples, as well as trace amounts of demethoxycurcumin. This suggests that *P. putida* might synthesize ferulic acid from coumaric acid, even without expressing *c3h* and *comt* genes, which can lead to the production of demethoxycurcumin and curcumin from this substrate. Furthermore, although FeCua1 expresses *curs1* codon optimized for

P. putida and FeCuaCg expresses *curs1* codon-optimized for *C. glutamicum*, they differed on the produced bisdemethoxycurcumin (0.01 ± 0.00 mM and 0.06 ± 0.01 mM, respectively), showcasing the effect of codon optimization on production (Table 2.2). The highest yield of p-coumaric acid was achieved by FeCua2 (24.0 ± 8.7%). Although this yield is below the reported 60% conversions by *E. coli* [115, 137], it represents a considerable improvement to previous attempts of bisdemethoxycurcumin production by *P. putida* that showed conversions below 1% [123].

Table 2.2: Production of curcumin from ferulic acid (FEA) or bisdemethoxycurcumin from coumaric acid (CUA). The mean and standard deviation of three replicates are shown. The yield is expressed as a percentage of the maximum theoretical yield. Production experiments were performed with an initial OD600 = 0.3 unless indicated by [a] (OD600 = 0.6) or [b] (OD600 = 0.9). p22 refers to the pSEVAb22 plasmid backbone.

Strain Name	Plasmids	Substrate (mM)	Titer (mM)	Yield (%)
FeCua1	p22-curs1-dcs	FEA (1 mM)	0.05 ± 0.02	9.73 ± 3.01
		FEA (2 mM)	0.17 ± 0.01	16.68 ± 0.98
		CUA (1 mM)	0.01 ± 0.00	4.90 ± 3.28
FeCua2	p22-curs2-dcs	FEA (1 mM)	0.16 ± 0.02	32.11 ± 3.34
		FEA (2 mM)	0.34 ± 0.05	34.37 ± 5.20
		FEA (2 mM)[a]	0.50 ± 0.02	50.00 ± 1.56
		FEA (2 mM)[b]	0.64 ± 0.03	63.60 ± 2.97
		CUA (1 mM)	0.12 ± 0.04	23.95 ± 8.73
FeCua3	p22-curs3-dcs	FEA (2 mM)	0.06 ± 0.00	6.19 ± 0.30
FeCuaCg	p22-cursCg-dcs	FEA (1 mM)	0.07 ± 0.01	14.01 ± 2.64
		FEA (2 mM)	0.18 ± 0.09	18.13 ± 8.89
		CUA (1 mM)	0.06 ± 0.01	13.03 ± 2.88
FeCuaCgCg	p22-cursCg-dcsCg	FEA (2 mM)	0.00 ± 0.00	0.34 ± 0.01

Production of curcuminoids from caffeic acid

After the successful synthesis of curcuminoids from ferulic and p-coumaric acid, the biosynthetic pathway for curcuminoid production was further expanded to produce curcumin from caffeic acid. This involved the introduction of two additional genes into a pSEVAb83 plasmid: *comt* and *ccoaomt* (Figure 2.2). Through the expression of both the pSEVAb22-curs-dcs and the pSEVAb83-comt-ccoaomt plasmids, five *P. putida* Δ*ech* strains able to synthesized curcumin from caffeic acid were constructed: Caa1, Caa2, Caa3, CaaCg, and CaaCgCg (Table 2.3).

Initially, the strains were tested with 2 mM caffeic acid. However, the curcumin levels were low (Caa1, Caa2, CaaCg) or zero (Caa3), and the lag phase of the cultures increased (Table 2.3, Sup. Figure 2.4), confirming the high toxicity of CAA found during the toxicity assay. Subsequently, Caa1, Caa2, and CaaCg, the strains with the best performance, were tested with 1 mM caffeic acid. The growth profile improved and the levels of curcumin significantly increased (Figure 2.2C,

Table 2.3, Sup. Figure 2.4). Similar to the experiments with ferulic and p-coumaric acids, curcumin degradation was observed after 70 h. Caa1 and CaaCg produced 0.17 ± 0.02 and 0.24 ± 0.05 mM of curcumin, respectively while Caa2 produced 0.10 ± 0.02 mM. Furthermore, ferulic acid was accumulated in large amounts for Caa1 (0.19 ± 0.04 mM) and Caa2 (0.57 ± 0.02 mM). The highest caffeic acid conversion, achieved by CaaCg (48.47 ± 9.06%), was 22.8 times higher than that obtained using *E. coli* [116].

Table 2.3: Production of curcumin from caffeic acid (CAA). The mean and standard deviation of three replicates are shown. The yield is expressed as a percentage of the maximum theoretical yield. p22 and p83 refer to the pSEVAb22 and pSEVAb83 plasmid backbones, respectively.

Name	Plasmids	CAA (mM)	Titer (mM)	Yield (%)
Caa1	p22-curs1-dcs + p83-comt-ccoaomt	1	0.17 ± 0.02	35.53 ±4.56
		2	0.00 ± 0.00	7.79 ± 0.45
Caa2	p22-curs2-dcs + p83-comt-ccoaomt	1	0.10 ± 0.02	19.27 ± 3.38
		2	0.00 ± 0.00	0.13 ± 0.03
Caa3	p22-curs3-dcs + p83-comt-ccoaomt	2	0.00 ± 0.00	0.15 ± 0.04
CaaCg	p22-cursCg-dcs + p83-comt-ccoaomt	1	0.24 ± 0.05	48.47 ± 9.06
		2	0.14 ± 0.01	14.84 ± 1.33
CaaCgCg	p22-cursCg-dcsCg + p83-comt-ccoaomt	2	0.00 ± 0.00	0.11 ± 0.19

Production of curcuminoids from tyrosine

After successfully producing curcumin from caffeic acid, the curcuminoids pathway was expanded to include heterologous expression of *tal* and *c3h* (Figure 2.2). Four strains able to produce curcuminoids from tyrosine were built differing in the *curs* isoenzyme used and the codon optimization: Tyr1, Tyr2, Tyr3, and TyrCg (Table 2.4). These strains carried pSEVAb22-curs-dcs and pSEVAb83-comt-ccoaomt-tal-c3h plasmids.

Considering the low tyrosine toxicity and the expected lower curcuminoid production from this substrate [114, 115, 116, 138], 3 mM tyrosine was used as the initial substrate concentration. However, low curcuminoid concentrations were found for Tyr1, Tyr2, and TyrCg and no production was observed for Tyr3 (Table 2.4, Sup. Figure 2.5). Therefore, a new experiment with 1 mM of tyrosine was performed using the producing strains Tyr1, Tyr2, and TyrCg. Reducing the concentration of tyrosine reduced the lag phase of the cultures and improved curcuminoid production (Figure 2.2A, Table 2.4, Sup. Figure 2.5). The highest production was achieved by Tyr2 (0.013 ± 0.001 mM) followed by Tyr1 (0.010 ± 0.005 mM) and TyrCg (0.005 ± 0.001 mM). Notably, although production of all curcuminoids was possible, curcumin was the only curcuminoid measured. The highest curcumin yield achieved (2.7 ± 0.2% of the maximum theoretical yield) was higher than previously reported conversions achieved with *E. coli* [114]. As observed with the other substrates, curcumin concentration decreased after 70 h indicating the degradation of this compound (Sup. Figure 2.5).

Additionally, Tyr1, Tyr2, Tyr3, and TyrCg were grown with glucose as only substrate to study curcuminoid production from endogenous tyrosine levels. Although no curcuminoids were detected in the Tyr3 strain, low curcumin concentrations were detected in cultures with Tyr1, Tyr2, and TyrCg (Sup. Figure 2.5), demonstrating curcuminoid production from glucose by *P. putida* Δ*ech* expressing the complete curcuminoid pathway.

Table 2.4: Production of curcumin from tyrosine (TYR). The mean and standard deviation of three replicates are shown. The yield is expressed as a percentage of the maximum theoretical yield. p22 and p83 refer to the pSEVAb22 and pSEVAb83 plasmid backbones, respectively.

Name	Plasmids	TYR (mM)	Titer (mM)	Yield (%)
Tyr1	p22-curs1-dcs + p83-comt-ccoaomt-tal-c3h	1	0.01 ± 0.00	1.30 ± 0.01
		3	0.00 ± 0.00	0.31 ± 0.31
Tyr2	p22-curs2-dcs + p83-comt-ccoaomt-tal-c3h	1	0.01 ± 0.00	2.67 ± 0.21
		3	0.01 ± 0.00	0.71 ± 0.21
Tyr3	p22-curs3-dcs + p83-comt-ccoaomt-tal-c3h	3	0.00 ± 0.00	0.00 ± 0.00
TyrCg	p22-cursCg-dcs + p83-comt-ccoaomt-tal-c3h	1	0.01 ± 0.00	1.11 ± 0.16
		3	0.00 ± 0.00	0.10 ± 0.17

Model-based optimization of curcumin production from tyrosine

Although production of curcuminoids from ferulic acid, p-coumaric acid, caffeic acid, and tyrosine was achieved, maximum theoretical yields were not obtained, indicating the possibility of further pathway optimization. To facilitate this optimization, we developed kinetic models of the curcuminoid pathway, allowing to monitor the concentrations of biomass, substrate, products, and intermediates over time. These models enabled the calculation of reaction fluxes and the impact of enzyme levels on production. Consequently, the insights gained through simulations were leveraged to steer the pathway optimization process. Considering that the lowest measured production was achieved with tyrosine as substrate and that curcumin was the main curcuminoid found, the optimization strategy focused on improving curcumin production from tyrosine.

Construction of kinetic models of the curcuminoid pathway

Kinetic models of the curcuminoid pathways expressed in FeCua2, Caa2, and Tyr2 strains were created. Considering that these models shared some reactions and their corresponding parameters (Table 2.1, Figure 2.3A-C), experimental data obtained with all these strains was simultaneously used for the parameterization of the models. Although this approach aimed to facilitate parameter estimation including experiments performed with different strains and substrates, it was not sufficient to obtain accurate parameter estimates. Instead, ensemble models formed by the 10 estimated parameter sets with the best agreement to experimental data (lowest least squares) were created and used for the simulations (enFeCua2, enCaa2, enTyr2).

Figure 2.3D shows the fit of the ensemble models to the experiments used during parameter estimation. All models accurately described precursor profiles except enFeCua2 which incorrectly simulated ferulic acid accumulation with ferulic acid as substrate in order to correctly fit the curcumin production curve. All models accurately simulated curcumin and bisdemethoxycurcumin production with normalized mean square errors (nMSE) of 2.2% and 3.5%. Demethoxycurcumin could only be produced by the Tyr2 strain, used in one of the four experiments for model training, which resulted in a higher nMSE (23.2%) when simulating the production of this metabolite. Besides, model simulations pointed at the accumulation of feruloyl-CoA, coumaroyl-CoA, and, to a lesser extent, caffeoyl-CoA, and diketide forms as causes for curcuminoid yields below 100% (Sup. Figure 2.6).

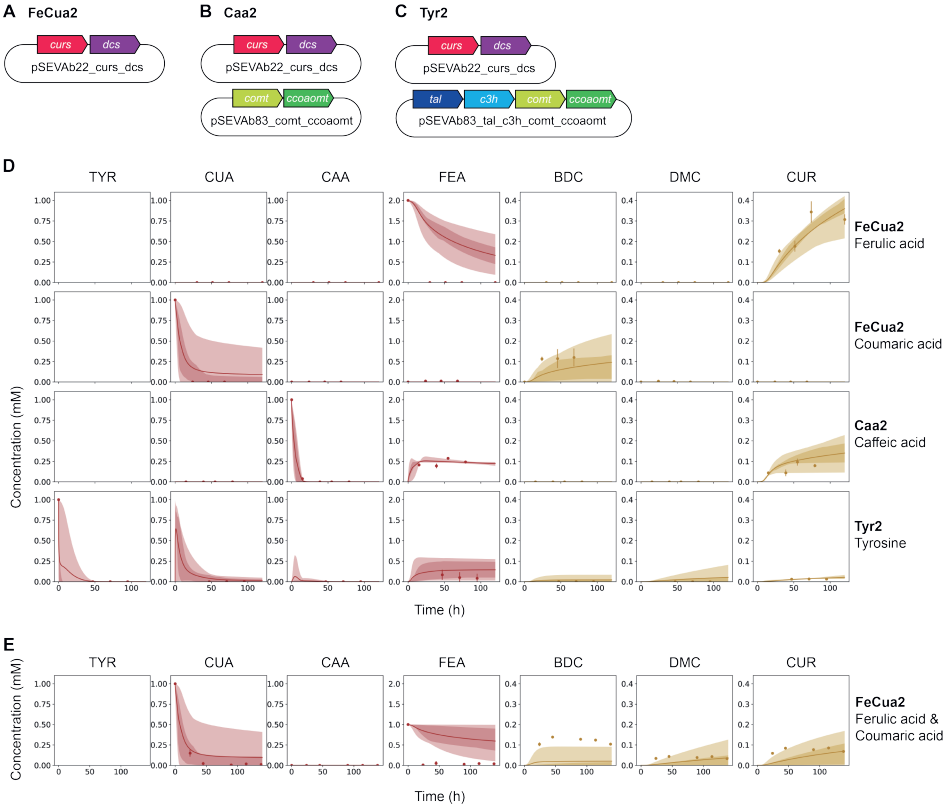


Figure 2.3: Comparison of ensemble model simulations and experimental data. **A-C.** Plasmids carried by the strains used for parameter estimation (FeCua2, Caa2, and Tyr2). **D.** Fit of ensemble models to the experimental data used for parameter estimation, the strains and substrate used are indicated. **E.** Fit of ensemble models to experimental data of validation experiment, the strain and substrates used are indicated. Points indicate experimental data, continuous lines represent the mean of the ensemble model predictions and shaded areas indicate 95-5% and 75-25% confidence intervals.

The ensemble modeling approach was validated simulating the performance of the FeCua2 strain simultaneously using ferulic and p-coumaric acids as substrates. Model predictions of curcuminoids showed good agreement with experimental data with a nMSE of 1.9%, 8.6%, and 1.2% for CUR, BDCUR, and DCUR, respectively (Figure 2.3E). Even though the ensemble failed to predict the complete depletion of ferulic acid, it predicted curcumin as the main curcuminoid, capturing the reported preference of CURS2 towards feruloyl-CoA [117].

Alleviation of enzymatic bottlenecks

After validating the qualitative performance of the ensembles, the focus was set on optimizing curcumin production from tyrosine. Therefore, the enTyr2 ensemble model was used to calculate the maximum fluxes through each curcuminoid pathway reaction during production with tyrosine as substrate (Figure 2.4I). Fluxes through the last two steps of the pathway (DCS and CURS) and the FCS reaction consuming ferulic acid were lower than those from the upper reactions (TAL, C3H, COMT, CCOAOMT), which explained the accumulation of pathway intermediates (Sup. Figure 2.6). In the Tyr2 strain, *curs2* and *dcs* genes were expressed in a plasmid with a lower copy number compared to the other genes in the pathway, and *fcs* expression was subject to native regulation (Figure 2.4A). The expected lower concentration of CURS2 and DCS enzymes was consistent with the lower fluxes calculated by the ensemble.

The flux imbalance among pathway reactions was addressed in the Tyr2_Opt1 and Tyr2_Opt2 strains by introducing *curs2* and *dcs* in higher copy number plasmids. These strategies aimed to increase gene expression of *curs2* and *dcs* to obtain higher enzyme concentrations and possibly improve the flux through the associated reactions. While Tyr2_Opt1 carried *curs2* and *dcs* in the high-copy number backbone pSEVAb25 and *comt*, *ccoaomt*, *tal*, and *c3h* in the medium-copy number backbone pSEVAb83; Tyr2_Opt2 carried these genes in the medium-copy number plasmid pSEVAb83 and the low-copy number plasmid pSEVAb22 (Table 2.5, Figure 2.4B, C). Production experiments were performed using 1 mM tyrosine as substrate and resulted in decreased production of curcumin by Tyr2_Opt1 and increased production of Tyr2_Opt2 (Figure 2.4J). While Tyr2 only produced curcumin at a final concentration of 0.013 $\pm$ 0.001 mM, Tyr2_Opt2 produced 0.021 $\pm$ 0.003mM curcumin and 0.031 mM of total curcuminoids (Figure 2.4J). This represented a 1.6-fold increase in curcumin production, a 2.4-fold increase in curcuminoid production, and a curcuminoid yield of 5.7 $\pm$ 0.3 % of the theoretical maximum. The effect of changing plasmid copy number was also assessed in strains expressing *curs1* codon-optimized for *C. glutamicum* instead of *curs2*, and similar results were found (Sup. Data 2).

Enhanced curcuminoid production was achieved by increasing the copy number of *dcs* and *curs2*. However, this improvement was only observed when coupled with a reduced expression of the other pathways genes, underscoring that overall higher gene expression does not guarantee a higher production. Rather, carefully selecting the genes to over-express was required. To facilitate this task, we employed the enTyr2 ensemble model to systematically evaluate the impact of varying enzyme concentrations on curcumin production from tyrosine. For each enzyme in

the pathway, the effect of reducing and increasing its concentration was assessed simulating curcumin production with the enzyme concentration estimated for each model in the ensemble as reference (Figure 2.4K, Sup. Figure 2.7).

Increasing the concentration of C3H and decreasing the concentration of COMT had the biggest positive impact on curcumin concentration *in silico* (Figure 2.4K). Increasing the concentration of C3H facilitates the consumption of p-coumaric acid. In turn, decreasing the concentration of COMT limits the production of ferulic acid which is predicted to accumulate due to the lower consumption flux by FCS (Figure 2.4I). In order to test model predictions, strains Tyr2_Opt3, with higher expression of C3H, and Tyr2_Opt4, not expressing COMT, were constructed and tested with 2 mM of tyrosine (Table 2.5, Figure 2.4L). Tyr2_Opt3 produced 0.07 $\pm$ 0.01 mM of curcumin, improving curcumin production 1.3-fold compared to Tyr2_Opt2. However, Tyr2_Opt4 produced 24.4% less curcumin than Tyr2_Opt2. The decreased production observed in Tyr2_Opt4 was accompanied by the accumulation of caffeic and ferulic acids, which suggested the conversion of feruloyl-CoA to ferulic acid by FCS, a behavior only predicted by one of the ten models of the ensemble (Sup. Figure 2.8).

Table 2.5: Production of curcumin from tyrosine (TYR) of optimized strains. The mean and standard deviation of three replicates are shown. The yield is expressed as a percentage of the maximum theoretical yield. p22, p83, and p25 refer to the pSEVAb22, pSEVAb83, and pSEVAb25 plasmid backbones, respectively.

Name	Plasmids	TYR (mM)	Titer (mM)	Yield (%)
Tyr2_Opt1	p25-curs2-dcs + p83-comt-ccoaomt-tal-c3h	1	0.01 $\pm$ 0.00	1.67 $\pm$ 0.13
Tyr2_Opt2	p83-curs2-dcs + p22-comt-ccoaomt-tal-c3h	1	0.02 $\pm$ 0.00	4.19 $\pm$ 0.58
		2	0.06 $\pm$ 0.00	5.68 $\pm$ 0.33
Tyr2_Opt3	p83-curs2-dcs-c3h + p22-comt-ccoaomt-tal	2	0.07 $\pm$ 0.01	6.97 $\pm$ 0.46
Tyr2_Opt4	p83-curs2-dcs + p22-ccoaomt-tal-c3h	2	0.03 $\pm$ 0.01	3.13 $\pm$ 0.66
Tyr2_Opt5	p83-curs2-dcs + p22-comt-ccoaomt-tal-c3h-4cl4	2	0.08 $\pm$ 0.01	8.22 $\pm$ 0.57
Tyr2_Opt6	p83-curs2-dcs-4cl4 + p22-comt-ccoaomt-tal-c3h	2	0.03 $\pm$ 0.00	3.07 $\pm$ 0.30
Tyr2_Opt7	p83-curs2-dcs-c3h + p22-comt-ccoaomt-tal-4cl4	2	0.11 $\pm$ 0.02	10.82 $\pm$ 1.81

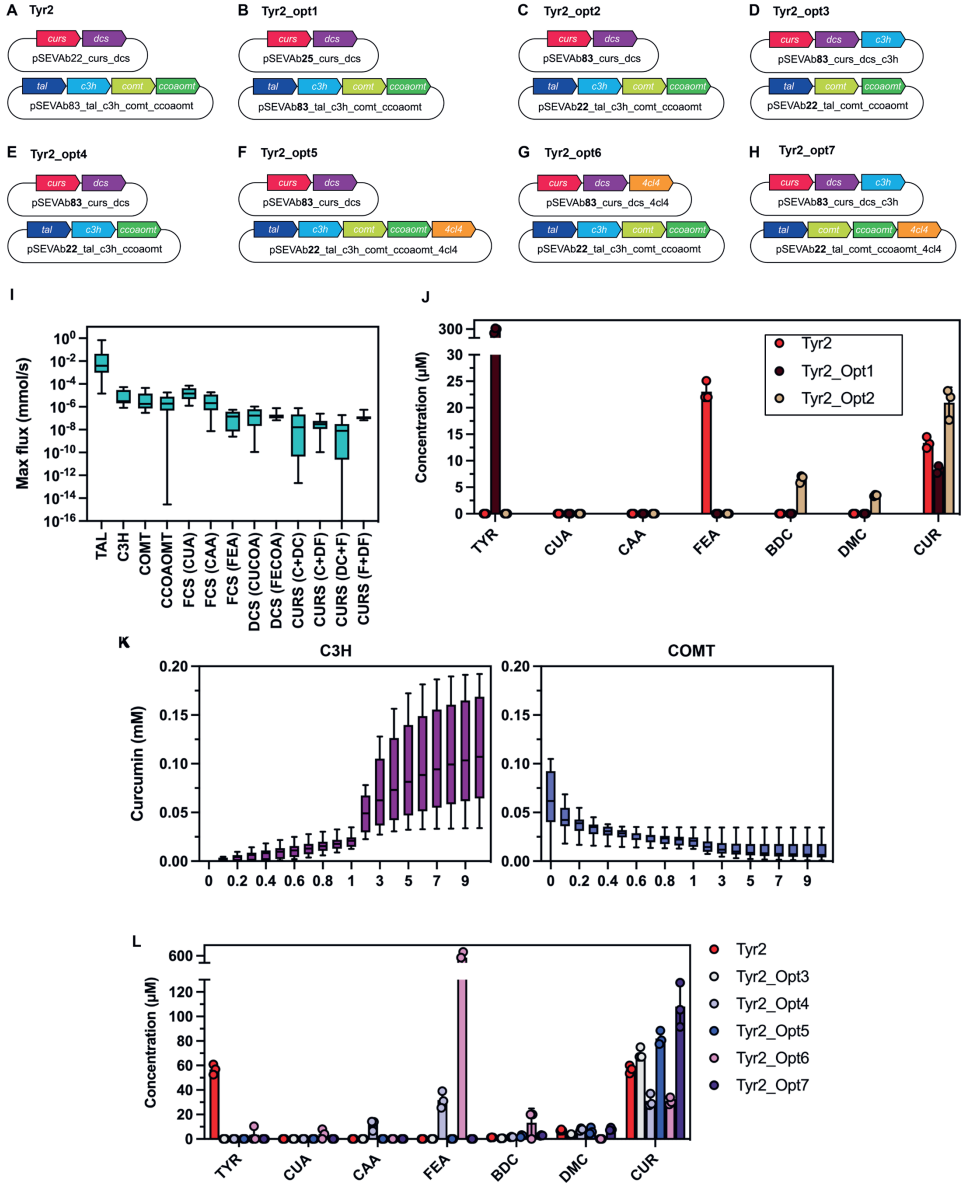


Figure 2.4: Model-based optimization of curcumin production from tyrosine. **A-H.** Plasmids carried by the optimized strains. **I.** enTyr2 ensemble predictions of maximum fluxes of each curcuminoid pathway reaction with tyrosine as substrate. For promiscuous enzymes, the substrates of each reaction are indicated between brackets: p-coumaric acid (CUA), caffeic acid (CAA), ferulic acid (FEA), coumaroyl-CoA (CUCOA), feruloyl-CoA (FECOA), diketyde coumaroyl-CoA (DC), diketyde feruloyl-CoA (DF). **J**. Effect of changing relative expression of *dcs* and *curs2* in Tyr2_Opt strains. **K.** enTyr2 ensemble predictions of curcumin production from tyrosine as a function of relative C3H and COMT enzyme concentrations. **L.** Effect of changing relative expression of *c3h* and *4cl4*, and omitting the expression of *comt* on curcumin production from tyrosine in Tyr2_Opt strains.

Expression of 4CL4 improves curcumin production

FCS was the only endogenous enzyme used in the curcuminoid pathway. Although this enzyme can convert p-coumaric, caffeic, and ferulic acid into their CoA-forms, kinetic information regarding substrate preference was not available. Besides, endogenous regulation of *fcs* expression in response to metabolites in the curcuminoid pathway is unknown. Although FCS related parameters could not be estimated, model simulations and experimental data suggested a possible detoxifying role of FCS that could result in the conversion of caffeic acid into ferulic acid which negatively impacted curcuminoid production. To reduce the accumulation of ferulic acid, alternatives to *fcs* were sought. *C. longa* and *Arabidopsis thaliana* express 4CL enzymes that perform functions analogous to FCS. *A. thaliana*'s 4CL4 enzyme has a 7.6-fold lower Michaelis Menten constant for ferulic acid compared to the enzyme from *C. longa*. Therefore, this enzyme has a higher affinity to ferulic acid and was included in new versions of the pathway [139].

The role of 4CL4 to complement FCS *in vivo* was tested in three strains: Tyr2_Opt5, Tyr2_Opt6, and Tyr2_Opt7 (Table 2.5, Figure 2.4F-H). Tyr2_Opt5 and Tyr2_Opt6 were based on Tyr2_Opt2 and were used to test the effect of expressing *4cl4* in pSEVAb22 and pSEVAb83 respectively. While expression of *4cl4* in the low-copy number plasmid improved curcumin production up to 0.08 $\pm$ 0.01 mM, expression in the medium-copy number plasmid reduced production to 0.03 $\pm$ 0.00 mM. When the expression of *4cl4* in pSEVAb22 was combined with the expression of *c3h* in pSEVAb83 in the Tyr2_Opt7 strain, 0.11 $\pm$ 0.02 mM of curcumin was achieved, a 1.6-fold increase compared to Tyr2_Opt3 and a 4-fold increase compared to Tyr2 (Figure 2.4L).

Discussion

In this study, we demonstrated the potential of using *P. putida* Δ*ech* as a platform for the bio-based production of curcuminoids, especially curcumin. We followed a systematic metabolic engineering framework, progressively expanding the substrate repertoire for curcuminoid production. Each step introduced new genes, increasing the pathway's complexity. Additionally, we evaluated, for the first time, the impact of different *curs* isoenzymes on curcuminoid production, achieving the highest reported conversion of caffeic acid and tyrosine to curcumin when utilizing *cursCg* and *curs2,* respectively. The acquired data confirmed the importance of the pathway kinetics on production and served as the foundation for training kinetic models of the curcuminoid pathway. These models enhanced the interpretation of the conducted experiments, proposed the development of new strains, and guided the hypothesis generation process during pathway optimization. This methodology culminated in a 4-fold enhancement in tyrosine conversion to curcumin, increasing the yield from 2.7 $\pm$ 0.2% to 10.8 $\pm$ 1.8% of the theoretical maximum, an 8.5-fold improvement over previously reported conversions [114].

Although CURS1, CURS2, and CURS3 have high amino acid identity, their kinetic parameters are considerably different [117]. While CURS1 has the lowest affinity (k_M) for ferulic and p-coumaric acid-derived molecules, it has the highest turnover rate (k_{cat}). In contrast, CURS3 shows the high-

est affinity towards these substrates but the lowest turnover rate. Possibly, the combination of high affinity of CURS2 and its turnover rate (twice higher than CURS3) is responsible for the better conversions shown in this study. Moreover, the preference of the FeCua2 strain for the ferulic acid over p-coumaric acid conversion is also explained by the 20-fold lower k_M constant of this enzyme to feruloyl-CoA compared to coumaroyl-CoA [117]. Notably, despite the consistent accumulation of ferulic acid when caffeic acid is used as a substrate, strains expressing *curs1* codon-optimized for *P. putida* or *C. glutamicum* achieved higher conversion rates compared to strains expressing *curs2*. A higher toxicity of caffeic acid on the Caa2 strain was discarded as the cause of the lower conversion since the growth of this strain was comparable to the Caa1 and CaaCg strains.

The effect of CURS isoenzymes on production emphasized the importance of kinetic parameters for pathway optimization given the promiscuity of CURS, FCS, and DCS. To better understand the pathway's kinetics, we constructed kinetic models of the pathway. Although accurate parameter estimation was not achieved, we developed ensemble models for the FeCua2, Caa2, and Tyr2 strains. Flux analyses using the enTyr2 model, combined with simulations assessing the influence of enzyme concentration on production, guided the development of the Tyr2_Opt3 strain. Within this strain, the genes *dcs*, *curs2*, and *c3h* are expressed in the medium-copy number plasmid (pSEVAb83) while *comt*, *ccoaomt*, and *tal* are expressed in a low-copy number plasmid (pSEVAb22). This strain achieved a 7.0 $\pm$ 0.5% conversion of tyrosine to curcumin, a 2.6-fold improvement compared to Tyr2.

Despite the consistent accuracy of ensemble models in simulating curcuminoid production, they were unable to replicate the complete depletion of ferulic acid in the absence of caffeic acid or tyrosine. This disagreement between experimental data and simulations was key to identifying the need to complement the endogenous expression of *fcs* with the expression of *A. thaliana*'s *4cl4*. Maximum reaction fluxes calculated by enTyr2 revealed a low flux of the FCS reaction consuming ferulic acid compared to the FCS reactions consuming p-coumaric and caffeic acids (Figure 2.4I). Besides, model simulations suggested omitting the expression of *comt* to improve curcumin production (Figure 2.4K). These observations led to the hypothesis that, in the absence of caffeic acid, FCS can convert ferulic acid into its CoA form but, when caffeic acid is present, FCS saturates with p-coumaric and caffeic acids resulting in the accumulation of ferulic acid. This accumulation is accentuated by the direct conversion of caffeic acid into ferulic acid by COMT. Although model simulations suggested reducing *comt* expression to improve curcumin production, excluding this gene in the Tyr2_Opt4 strain decreased curcumin production and led to ferulic acid accumulation (Figure 2.4L). In this strain, FCS catalyzes the reverse conversion from feruloyl-CoA to ferulic acid, a mechanism only captured by one of the models in the ensemble (Sup. Figure 2.8). Therefore, we addressed the possible limitation caused by FCS by expressing *4cl4* in the Tyr2_Opt7 strain. As a result, we achieved a conversion efficiency of 10.8 $\pm$ 1.8% from tyrosine to curcumin, 1.6-fold and 4-fold improvement compared to Tyr2_Opt3 and Tyr2, respectively. In addition, we showed the ability of the ensemble modeling approach to study the curcuminoid pathway even when accurate parameter values could not be estimated.

Beyond gene expression, the observed degradation of curcumin, the substrate concentration, and the initial OD600 of the culture influenced the curcumin yields. While photodegradation of curcuminoids can be prevented by shielding the cultivation medium from light exposure [106], the NADPH-dependent curcumin/dihydrocurcumin reductase (CurA) can catalyze the breakdown of curcumin. The deletion of the *curA* gene improved curcumin yield in *E. coli* [140] and could be implemented in *P. putida* Δ*ech*. Furthermore, the initial substrate concentration was a major determinant in curcuminoid production outcomes. The optimum substrate concentration varied depending on the strain, with strains exhibiting superior conversion capabilities tolerating higher substrate concentrations. Including the effect of substrate and intermediate concentrations on growth in the ensemble models could therefore be used to estimate optimum initial concentrations and further enhance production. Alongside substrate concentration, other process-related parameters can be tuned to improve production [141]. For example, we showed how increasing the initial OD600 of the cultures resulted in a 1.8-fold increase in curcumin production from ferulic acid (Table 2.2, Sup. Figure 2.2).

P. putida Δ*ech* showed lower conversion yields when using ferulic and p-coumaric acids as substrates compared to *E. coli* [114, 115]. However, as the complexity of the pathway increased, and caffeic acid and tyrosine were used as substrates, *P. putida* Δ*ech* strains out-competed previously reported *E. coli* strains [114, 116]. The 10-fold increase in caffeic acid conversion by CaaCg, Caa1, and Caa2 compared to *E. coli* could be attributed to the expression of *comt* or the higher tolerance of *P. putida* to this substrate. Although conversions up to 11.5% have been reported with tyrosine as substrate, they resulted in the production of bisdemethoxycurcumin [115, 138]. These strains were unable to produce caffeic and ferulic acids which reduced the number of possible toxic pathway metabolites and likely favored production. When *c3h* and *comt* were expressed in one of these strains, the production of curcumin was reduced to 0.1% of the maximum theoretical yield [138]. This yield was improved by Rodrigues et al. to 1.3% expressing *dcs* and *curs1* instead of *cus* [114], probably due to the higher affinity of *curs1* to ferulic acid-derived molecules compared to *cus* [142]. Higher yields of curcumin from tyrosine (2.9% of the theoretical maximum) have only been achieved by the use of a co-culture with two *E. coli* strains, each expressing a part of the pathway [114]. Notably, the Tyr2 strain alone matched the yield achieved with the co-culture, and Tyr2_Opt7 showed an 8.5-fold improvement, which highlights the potential of *P. putida* for curcumin production.

In summary, we established the production of curcuminoids in *P. putida*, optimized production yields from ferulic acid, p-coumaric acid, caffeic acid, and tyrosine, and provided the basis to produce curcuminoids from glucose. This was achieved by the creation of ensemble dynamic models to understand pathway kinetics and identify bottlenecks, the testing of various isoenzymes for their efficiency, and the fine-tuning gene expression levels and substrate concentrations for optimal yields. This comprehensive approach led to significant improvements in curcuminoid production, culminating in the highest yield (10.8 $\pm$ 1.8% of the theoretical maximum) of curcumin production from tyrosine reported to date.

Declaration of interest

Vitor A. P. Martins dos Santos has interests in LifeGlimmer GmbH and Richard van Kranenburg is employed by Corbion N.V.

Acknowledgment

This project was founded by NWO (project numbers GSGT.2019.008 and GSGT.2019.028). The authors would also like to acknowledge Silvia Rodriguez Marcos for help in constructing some of the strains in this study.

Data availability

Scripts, models, and supplementary tables are available at Gitlab and Zenodo.

Supplementary methods: curcuminoid pathway model

Three curcuminoid pathway models of different complexities were used to simulate the constructed strains. Summaries of the concentrations of metabolites tracked (states) and fluxes present in each model are shown in Sup. Methods Tables 2.1 and 2.2.

Ordinary differential equations (ODE) were used to simulate the time evolution of the state variables. Changes in growth were represented using a logistic equation:

$$\frac{dBiomass}{dt} = v_{growth} \tag{2.1}$$

$$v_{growth} = Biomass * k_{growth} * (1 - \frac{Biomass}{B_{max}}), \tag{2.2}$$

where k_{growth} is the growth constant and B_{max} the maximum biomass concentration supported by the medium.

Changes in metabolite concentrations were determined using mass balances, summing reaction rates of reactions producing the metabolites, and subtracting rates of reactions consuming the metabolite. For instance, for caffeic acid (CAA) concentration, the following equation was used:

$$\frac{dCAA}{dt} = v_{C3H} - v_{COMT} - v_{FCS\_CAA}, \tag{2.3}$$

where v_{C3H} represents the rates of the caffeic acid producing reaction C3H, and v_{COMT} and c_{FCS_CAA} are the rates of the COMT and FCS reactions consuming caffeic acid.

For all reactions but C3H, the generalized reversible Michaelis-Menten kinetics expression rate law was used [127]. For a reaction R converting metabolite A into B using cofactor X that gets converted into Y, this approach results in:

$$v_R = Biomass \cdot u_R^v \cdot \frac{A \cdot c_X \cdot e^{\frac{\mu_A + \mu_X - \mu_B - \mu_Y}{2RT}} - B \cdot c_Y \cdot e^{\frac{-\mu_A - \mu_X + \mu_B + \mu_Y}{2RT}}}{\sqrt{k_{A,R}^M \cdot k_{X,R}^M \cdot k_{B,R}^M \cdot k_{Y,R}^M} \left[\frac{A}{k_{A,R}^M + 1} \frac{X}{k_{X,R}^M + 1} + \frac{B}{k_{B,R}^M + 1} \frac{Y}{k_{Y,R}^M + 1} - 1 \right]}. \tag{2.4}$$

Here u_R^v represents the product of the enzyme concentration (u_R) and the catalytic rate constant (k_R^v) for reaction R; μ_i denotes the chemical potential of metabolite i, and $k_{i,R}^M$ its Michaelis-Menten constant. A and B represent the concentrations of the states (*i.e.* metabolites assumed to vary in time), whereas c_X and c_Y represent the concentration of the corresponding cofactors, which are assumed to remain constant. The gas constant and temperature are denoted as R and T, respectively. Rates and fluxes are scaled using the $Biomass$ concentration. For enzymes catalyzing multiple reactions (FCL, DCS, CURS), u_R and k_R^v parameters are considered separately and additional substrate competition terms including the Michaelis-Menten constant of all possible products and substrates are added to the specific reaction rate laws as specified by Liebermeister et al. [127].

Flux through the C3H reaction was modeled using mass action kinetics and scaled with the $Biomass$ concentration:

$$v_{C3H} = Biomass * u^v_{C3H} * CUA,$$

where u^v_{C3H} is the product of the enzyme concentration (u_R) and the catalytic rate constant (k^v_R) for the C3H reaction, and CUA is the concentration of p-coumaric acid, the only substrate of this enzyme.

A detailed description of the genes, reactions and metabolites included in the different models is available in Gitlab. Initial estimates used for metabolite and cofactor concentrations and kinetic parameters with their sources, as well as initial estimates for chemical potentials obtained using Equilibrator are provided. SBML versions of the three models are available in Gitlab.

Sup. Methods Table 2.1: Summary of metabolites considered as states in the different models. A state is considered an observable when it is experimentally measured.

State	Abreviation	Observable?	FeCua model	Caa model	Tyr model
Biomass	Biomass	Yes	x	x	x
Ferulic acid	FEA	Yes	x	x	x
Coumaric acid	CUA	Yes	x	x	x
Caffeic acid	CAA	Yes	x	x	x
Tyrosine	TYR	Yes			x
Feruloyl-CoA	FECOA	No	x	x	x
Coumaroyl-CoA	CUCOA	No	x	x	x
Caffeoyl-CoA	CACOA	No	x	x	x
Diketide feruloyl-CoA	DFECOA	No	x	x	x
Diketide coumaroyl-CoA	DCUCOA	No	x	x	x
Bis-demethoxycurcumin	BDCUR	Yes	x	x	x
Demethoxycurcumin	DCUR	Yes	x	x	x
Curcumin	CUR	Yes	x	x	x

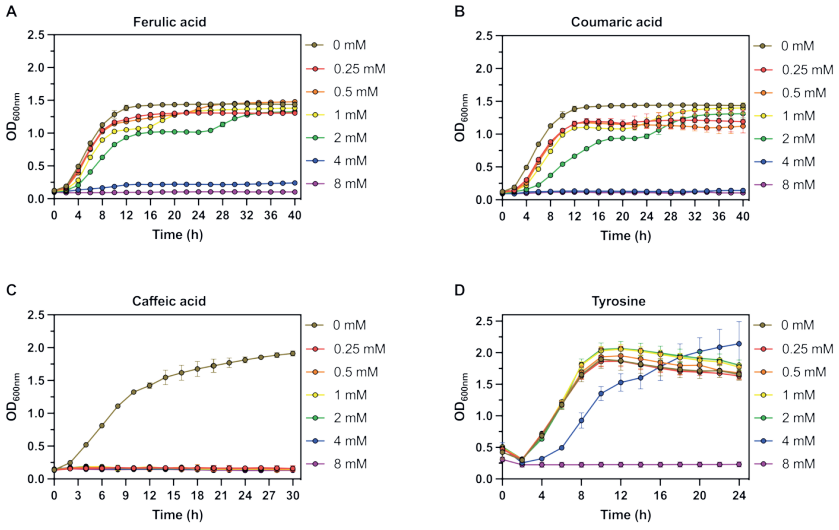
Sup. Methods Table 2.2: Summary of reactions included in the different models. Equation abbreviations: LOG, logistic equation; GMM, generalized Michaelis-Menten equation; MAK, mass action kinetics. Metabolite abbreviations: TYR, tyrosine; CUA, p-coumaric acid; CAA, caffeic acid; AMET, S-adenosyl methionine; CACOA, caffeoyl-CoA; FEA, ferulic acid; CUCOA, coumaroyl-CoA; MACOA, malonyl-CoA; FECOA, feruloyl-CoA; DCUCOA, diketide coumaroyl-CoA; DFECOA, diketide feruloyl-CoA; AHCYS, homocysteine; PPi, inorganic pirophosphate; BDCUR, bisdemethoxycurcumin; DCUR, demethoxycurcumin; CUR, curcumin.

Reaction	Equation	Substrates	Products	Models
Growth	LOG	-	-	FeCua, Caa, Tyr
TAL_TYR	GMM	TYR	CUA, NH_3	Tyr
C3H_CUA	MAK	CUA	CAA	Tyr
COMT	GMM	CAA, AMET	FEA, AHCYS, H	Caa, Tyr
CCOAOMT	GMM	CACOA, AMET	FECOA, AHCYS, H	Caa, Tyr
FCL_CUA	GMM	CUA, ATP, COA	CUCOA, AMP, PPi, H	FeCua, Caa, Tyr
FCL_CAA	GMM	CAA, ATP, COA	CACOA, AMP, PPi, H	FeCua, Caa, Tyr
FCL_FEA	GMM	FEA, ATP, COA	FECOA, AMP, PPi, H	FeCua, Caa, Tyr
DCS_CUCOA	GMM	CUCOA, MACOA, H_2O	DCUCOA, HCO_3, COA, H	FeCua, Caa, Tyr
DCS_FECOA	GMM	FECOA, MACOA, H_2O	DFECOA, HCO_3, COA, H	FeCua, Caa, Tyr
CURS_BDCUR	GMM	CUCOA, DCUCOA, H_2O	BDCUR, COA, HCO_3, H	FeCua, Caa, Tyr
CURS_DCUR1	GMM	CUCOA, DFECOA, H_2O	DCUR, COA, HCO_3, H	FeCua, Caa, Tyr
CURS_DCUR2	GMM	FECOA, DCUCOA, H_2O	DCUR, COA, HCO_3, H	FeCua, Caa, Tyr
CURS_CUR	GMM	FECOA, DFECOA, H_2O	CUR, COA, HCO_3, H	FeCua, Caa, Tyr

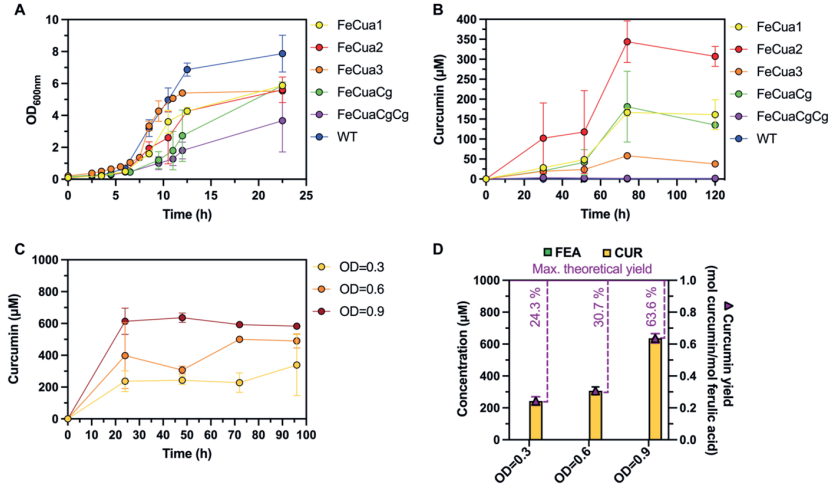
Supplementary Tables and Figures

Sup. Table 2.1: Summary of microbial curcuminoids production from different substrates (Sub.). FEA, ferulic acid; CUA; p-coumaric acid; CAA, caffeic acid; TYR, tyrosine; CUR, curcumin; BDC, bisdemethoxycurcumin. The yield is expressed as a percentage of the maximum theoretical yield.

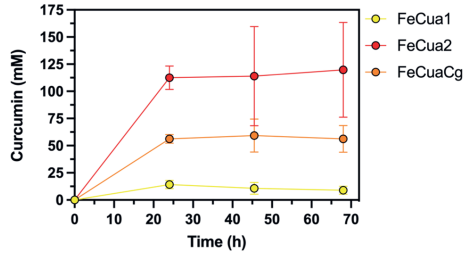
Sub.	Organism	Over-expressed Genes	%-Yield	Ref.
FEA	*E. coli* (1 mM)	*4cl*, *accABCD*, *cus*	CUR (61%)	[115]
	E. coli (2 mM)	*4cl*, *dcs*, *curs1*	CUR (96%)	[141]
	E. coli (2 mM)	*dcs*, *curs1*, *tal*, *c3h*, *ccoaomt*, *4cl*	CUR (19%)	[116]
	E. coli (2 mM)	*tal*, *c3h*, *4cl1*, *dcs*, *curs1*	CUR (91.7%)	[143]
	E. coli (3 mM)	*tal*, *c3h*, *4cl*, *comt*, *dcs*, *curs1*	CUR (100%)	[114]
	E. coli (2 mM)	*cus*, *4cl*	CUR (1.5%)	[144]
	E. coli Δ*curA*(4 mM)	*cus*, *4cl*, improve membrane & MalCoA	CUR (73%)	[144]
	S. cerevisiae (0.08 mM)	*4cl* , *ferA*, *cus*, *dcs*, *curs1*	CUR (17.8%)	[145]
CUA	*E. coli* (1 mM)	*4cl*, *accABCD*, *cus*	BDC(59%)	[115]
	E. coli (0.15 mM)	*cus*, *4cl*, *accABCD*, *matBC*, *tal*	BDC (58.6%)	[137]
	P. putida Δ*ech*(5 mM)	*cus*	BDC (0.2%)	[123]
	E. coli (2 mM)	*dcs*, *curs1*, *tal*, *c3h*, *ccoaomt*, *4cl*	BDC (0.08%)	[116]
	Y. lipolytica (2 mM)	*tal*, *cus*, *4cl*, *g2ps1*, *pks1*, *sts*, *bas*	BDC (0.05%)	[146]
CAA	*E. coli* (1 mM)	*dcs*, *curs1*, *tal*, *c3h*, *ccoaomt*, *4cl*	CUR (2.12%)	[116]
TYR	*E. coli* (3 mM)	*pal*, *4cl*, *accABCD*, *cus*	BCUR (11.5%)	[115]
	E. coli (3 mM)	*dcs*, *curs1*, *tal*, *c3h*, *ccoaomt*, *4cl*	CUR (0.04%)	[116]
	E. coli (3 mM)	*tal*, *c3h*, *4cl*, *comt*, *dcs*, *curs1*	CUR (1.27%)	[114]
	E. coli (3 mM)	*tal*, *4cl4*, *cus*	BCUR (6.2%)	[138]
	E. coli (3 mM)	*tal*, *4cl4*, *cus*, *c3h*, *comt*	CUR (0.1%)	[138]



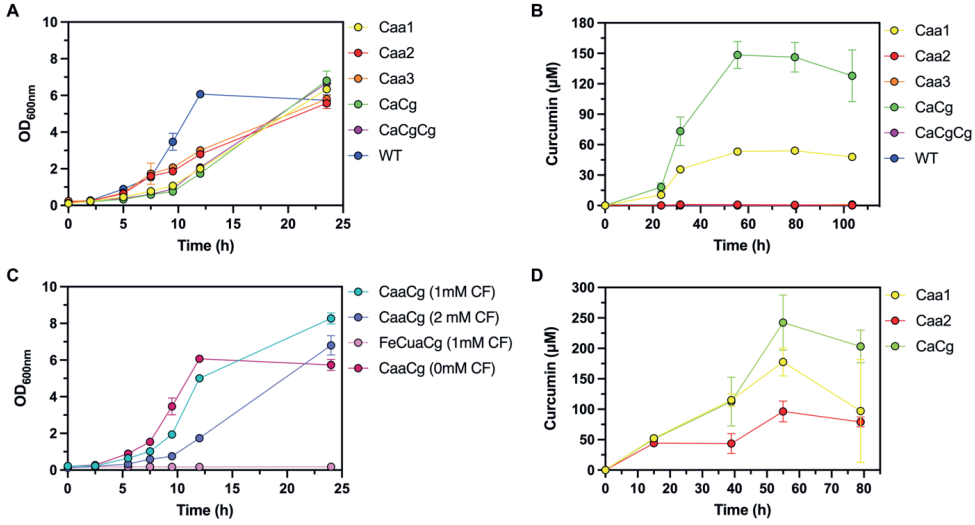
Sup. Figure 2.1: Evaluation of the toxicity of hydroxycinnamic acids and tyrosine on *P. putida* Δ*ech*. Growth curves of *P. putida* Δ*ech* in minimal M9 medium supplemented with 70 mM glucose and different concentrations of ferulic acid (**A**), coumaric acid (**B**), caffeic acid (**C**) and tyrosine (**D**). The OD600 of the cultures was measured over 24-40 hours at a wavelength of 600nm (OD600). Values represent the mean and the standard deviation of five biological replicates.



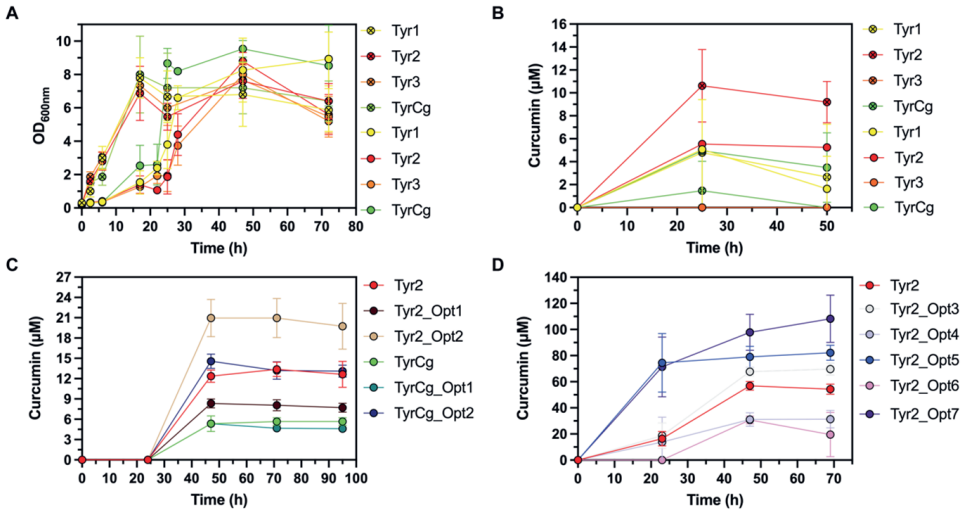
Sup. Figure 2.2: Production of curcumin using ferulic acid as substrate. Growth (**A**) and production (**B**) curves of different *P. putida* Δ*ech* strains in minimal M9 medium supplemented with 70 mM glucose and 2 mM ferulic acid. **C.** Production of curcumin of FeCua2 in minimal M9 medium supplemented with 70 mM glucose and 2 mM of ferulic acid over 96 h. 2 mM of ferulic acid were added at different OD600; 0.3, 0.6, 0.9. **D.** Ferulic acid and curcumin concentration of FeCua2 at 24 h when 2 mM of ferulic acid were added at different OD600; 0.3, 0.6, 0.9. Dotted purple lines indicate the achieved yields in the different conditions as a percentage of the maximum theoretical yield. Values represent the mean and the standard deviation of five biological replicates.



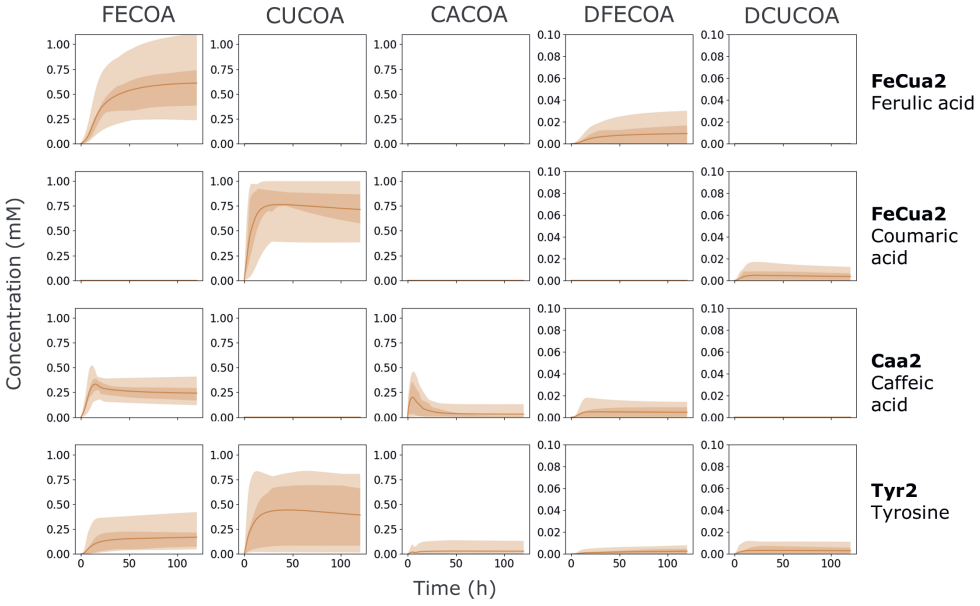
Sup. Figure 2.3: Production of bisdemethoxycurcumin over time by different *P. putida* Δ*ech* strains in minimal M9 medium supplemented with 70 mM glucose and 1 mM p-coumaric acid. Values represent the mean and the standard deviation of five biological replicates.



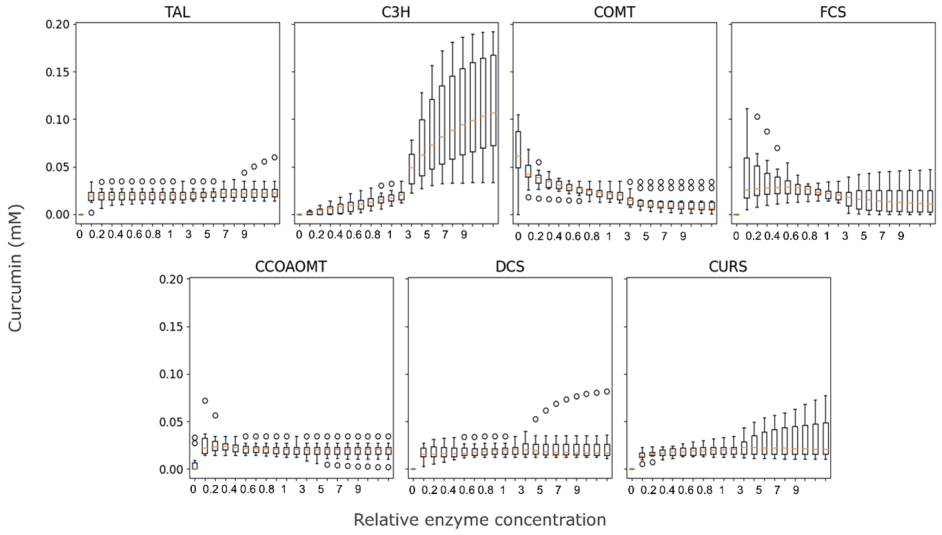
Sup. Figure 2.4: Production of curcumin using caffeic acid as substrate. Growth (**A**) and production (**B**) curves of different *P. putida* Δ*ech* strains in minimal M9 medium supplemented with 70 mM glucose and 2 mM caffeic acid. **C.** Growth curves of CaaCg and FeCuaCg strains measured as OD600 over time. **D.** Curcumin production of Caa1, Caa2, and Caa3 grown in minimal M9 medium supplemented with 70mM glucose and 1 mM of caffeic acid over 80 h. Values represent the mean and the standard deviation of five biological replicates.



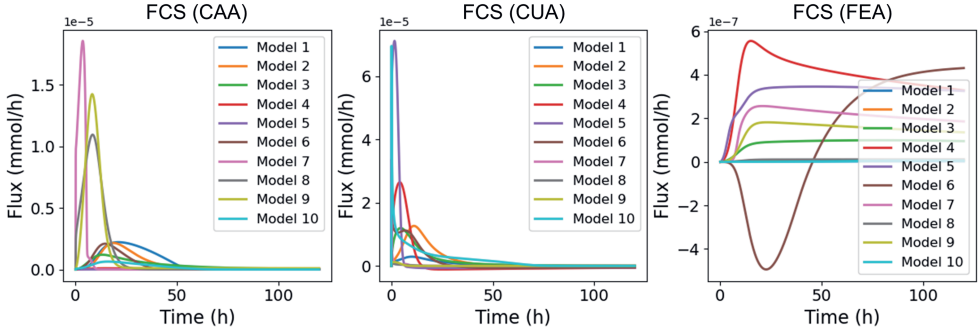
Sup. Figure 2.5: Production of curcumin using tyrosine as substrate. Growth (**A**) and production (**B**) curves of different *P. putida* Δ*ech* strains in minimal M9 medium supplemented with 70 mM glucose, 0 mM (empty circles), or 3 mM tyrosine (crossed circles). Production of curcumin of different Tyr2 strains using 1 mM (**C**) and 2 mM (**D**) of tyrosine. Values represent the mean and the standard deviation of five biological replicates.



Sup. Figure 2.6: Production of intermediates predicted by ensemble models using ferulic acid, coumaric acid, caffeic acid, or tyrosine as substrates. Continuous lines represent the mean of the ensemble model predictions and shaded areas indicate 95-5% and 75-25% confidence intervals. FECOA, feruloyl-CoA; CUCOA, coumaroyl-CoA; CACOA, caffeoyl-CoA; DFECOA, diketide-feruloyl-CoA; DCUCOA, diketide-coumaroyl-CoA.



Sup. Figure 2.7: enTyr2 ensemble predictions of curcumin production from tyrosine based on relative enzyme concentrations.



Sup. Figure 2.8: Predicted fluxes through reactions performed by FCS with tyrosine as substrate. Fluxes for each model in the ensemble carried by the FCS reactions consuming cafferic acid (CAA), coumaric acid (CUA), and ferulic acid (FEA) are shown.

CHAPTER 3

CFSA: Comparative Flux Sampling Analysis as a guide for strain design

Sara Moreno Paz*, Rik P. van Rosmalen*, Z. Efsun Duman Özdamar, María Suárez Diez

*Contributed equally

This chapter is under review in *Metabolic Engineering Communications*

Abstract

Genome-scale metabolic models of microbial metabolism have extensively been used to guide the design of microbial cell factories, still, many of the available strain design algorithms often fail to produce a reduced list of targets for improved performance that can be implemented and validated in a step-wise manner. We present Comparative Flux Sampling Analysis (CFSA), a strain design method based on the extensive comparison of complete metabolic spaces corresponding to maximal or near-maximal growth and production phenotypes. The comparison is complemented by statistical analysis to identify reactions with altered flux that are suggested as targets for genetic interventions including up-regulations, down-regulations, and gene deletions. We applied CFSA to the production of lipids by *Cutaneotrichosporon oleaginosus* and naringenin by *Saccharomyces cerevisiae* identifying engineering targets in agreement with previous studies as well as new interventions. CFSA is an easy-to-use, robust method that suggests potential metabolic engineering targets for growth-uncoupled production that can be applied to the design of microbial cell factories.

Introduction

Microbial cell factories are microorganisms engineered for the production of bio-molecules that thrive on renewable carbon sources. Using these microorganisms, a broad range of bio-molecules can be produced from non-edible feedstocks such as recalcitrant biomass or industrial waste streams thereby providing sustainable replacements for production systems based on fossil fuels [147]. The design of microbial cell factories requires the choice of an appropriate host strain, and the selection of a suitable available pathway, the discovery of new pathways, or the design of synthetic pathways for new-to-nature compounds. Still, industrial feasibility requires extensive engineering to improve the performance of the cell factory [148].

GEnome-scale Metabolic models (GEM) are comprehensive representations of the cell's metabolism that allow the simulation of metabolic fluxes and the prediction of cellular phenotypes. They have largely been used to identify metabolic engineering targets to optimize pathway performance [57]. Machado et al. classify strain design algorithms relying on GEMs into two main groups: methods based on the analysis of elementary flux modes (EFM), and those based on the optimization of an objective function [149].

EFMs are minimal sets of reactions that can jointly operate at steady-state such that all steady-state solutions can be described as a combination of EFMs [150]. They provide an unbiased framework to explore the metabolic space but have limited scaling potential and applicability to larger models. Instead, related approaches, based on minimal cut sets (MCS) such as minimal metabolic functionality (MMF) and FluxDesign, scale better to genome scale [149, 151].

Optimization-based approaches rely on the simulation of metabolic fluxes of *wild type* and/or mutant strains using an objective function. Many of these methods use Flux Balance Analysis (FBA) to calculate fluxes and therefore only explore one of the multiple flux distributions that can lead to the optimal objective ignoring the rest. In this way, OptKnock [152] and derived algorithms such as RobustKnock [153], OptGene [154], and OptCouple [155], aim at identifying gene knock-outs to couple the production of the compound of interest to the production of biomass using growth as the objective function. Other methods such as OptForce [156] and OptDesign [157] are used to predict modulation of gene expression, including up and down-regulations. They compare simulated fluxes of growth and production phenotypes and minimize the number of interventions required for the overproduction phenotype. Although flux variability analysis (FVA) might be used to explore feasible flux ranges and constrain the solution space, comparisons among phenotypes are based on non-unique FBA solutions. Alternatively, flux sampling allows the exploration of the full space of feasible flux distribution of each reaction given a set of constraints on the metabolic model [64]. Contrarily to FBA, flux sampling does not require the selection of an objective function, which biases FBA predictions. The advantages of flux sampling have been exploited to evaluate metabolic flux differences between different conditions [158], but have not yet been expanded to strain design.

The described strain design algorithms focus on the identification of growth-coupling strategies, where production becomes a requirement for growth [152, 153, 154, 155]. This strategy is suitable for experimental implementation and further optimization using adaptive laboratory evolution but requires multiple simultaneous interventions. Alternatively, strains with growth-uncoupled production can be used in two-stage fermentation processes where the growth and production phases are sequential. This strategy can alleviate metabolic stress and improve productivity [159, 160].

We present Comparative Flux Sampling Analysis (CFSA), a model-guided strain design approach based on extensive sampling of the feasible solution space in alternative scenarios. Growth and production phenotypes are simulated and compared, also with a growth-limited scenario, which serves as a negative control for down-regulation targets. Flux distributions are statistically compared resulting in the identification of potential over-expression, down-regulations, and knockout targets leading to growth-uncoupled increased production. As a proof of concept, we use CFSA to identify metabolic engineering targets for lipid production by *Cutaneotrichosporon oleaginosus* and naringenin production by *Saccharomyces cerevisiae* and compare them with available data.

Methods

Comparative Flux Sampling Analysis (CFSA)

Flux sampling

To implement CFSA a GEM of the desired organism including the production pathway of choice is required. As the first step, media conditions such as substrate uptake rates, or aerobic/anaerobic growth are specified. Model reactions are grouped into seven categories to facilitate later filtering: *required* reactions including the growth and maintenance reactions; *not biological* reactions including boundary, exchange, sink, and demand reactions; *blocked* reactions (unable to carry flux); reactions *without associated genes*; *essential* reactions; reactions *containing essential genes*; and *transport* reactions.

Reaction fluxes are sampled from the metabolic solution space in three scenarios: growth, slow growth, and production. In the growth and production scenarios, the optimality parameter ensures that sampled flux distributions result in at least a specified fraction of the optimal growth or production predicted by FBA by constraining the lower bound of the biomass or product exchange reactions. In the slow growth scenario, the maximum growth rate compatible with the specified minimal production rate is calculated and used as an upper bound for the biomass synthesis reaction. To limit the solution space a parsimonious FBA approach is implemented by introducing an additional constraint to limit the total sum of fluxes to the minimum value compatible with maximal growth given by the flux fraction parameter [161]. This extra constraint is applied in the production and slow growth scenario simulations to limit unrealistic futile cycles.

The Optimal Gaussian Process (OptGP) sampler implemented in cobrapy is used to model the distribution of the target space and iteratively sample from this distribution [162, 163]. A thinning parameter is used to reduce the correlation between samples. Invalid samples (*i.e.* those that do not meet constraints specific to each scenario) are discarded. Last, the Geweke diagnostic is used to calculate the chain convergence for each process, and samples corresponding to reactions whose distributions have not converged are discarded [164].

Filtering of metabolic targets

For each reaction, the two-sample Kolmogorov-Smirnov (KS) test is used to compare samples from different scenarios and determine if they belong to the same continuous distribution. Potential targets are selected if, for a specific reaction, distributions differ in the scenarios based on KS statistics and p-values corrected for multiple testing using Bonferroni. The p-value and KS cut-offs can be adjusted by the user. Besides, reactions whose fluxes correlate with fluxes through the biomass synthesis reaction, or that do not have a gene-protein-reaction association are discarded as potential targets. Only reactions whose absolute change in flux between the growth and production scenario is bigger than a user-specified threshold are considered to be suitable targets. Similarly, targets can be filtered based on the standard deviation of the samples taken in the production scenario.

Potential targets are then divided into over-expression or down-regulation targets depending on whether the mean fold change comparing growth and production scenarios is above or below one. Knock-down targets that correspond to non-essential genes are classified as possible knock-out targets. Reactions are clustered based on the correlation of the absolute fluxes between samples to identify redundant targets (*i.e.* belonging to the same metabolic pathway).

Selected applications

Production of lipids by *C. oleaginosus*

The *i*NP636_*Coleaginosus*_ATCC20509 genome-scale metabolic model was manually curated to provide all fatty acid elongation reactions [165]. In total seven reactions were added, one of them in the cytoplasm and six of them in the mitochondria, and all reactions were associated with the corresponding genes. Details of the model curation can be found in GitLab. Glycerol and urea were used as the carbon and nitrogen sources respectively and nitrogen-depleted biomass composition was assumed. CFSA was used (optimality = 0.90, flux fraction = 1.25, KS1 = KS2 $\geq$ 0.75, mean absolute change $\geq$ 0.01, standard deviation in production $\leq$ 50) using the lipid synthesis reaction (lipid_synthesis) as a target for the production scenario. The feasibility of selected targets was evaluated based on previous studies.

Production of naringenin by *S. cerevisiae*

The Yeast8 genome-scale metabolic model of *S. cerevisiae* [166] was modified to include the naringenin production pathway (Table 3.1) and glucose was the selected carbon source. CFSA was used (optimality = 0.90, flux fraction = 1.25, KS1 = KS2 $\geq$ 0.75, mean absolute change $\geq$ 0.01, standard deviation in production $\leq$ 50) with the naringenin exchange reaction (EX_NAR) as the objective for the production scenario. Two proteomic datasets representative of *S. cerevisiae* aerobic growth on glucose during shake flask and chemostat fermentations were used as additional filters to restrict down-regulation targets to detected proteins [167, 168].

Table 3.1: Naringenin pathway reactions added to Yeast8. Reaction names in the model, reaction equations, and Gene-Protein-Reaction (GPR) rules are included. PAL, phenylalanine ammonia-lyase; TAL, tyrosine ammonia-lyase; C4H, cinnamate 4-hydroxylase; CL, 4-coumarate-CoA ligase; CHS, naringenin-chalcone synthase; CHI, chalcone isomerase; NARt, naringenin transport; EX_NAR, naringenin exchange; L-Phe, L-phenylalanine; CIN, cinnamate; L-Tyr, L-tyrosine; 4-CUA, 4-coumarate; CoA, Coenzyme-A; CUACoA, 4-coumaroyl-CoA; MaCoA, malonyl-CoA; NAR-C, naringenin chalcone; NAR, naringenin. _c and _e designate metabolites in the cytoplasm and the extracellular space respectively.

Reaction name	Reaction equation	GPR
PAL	L-Phe_c $\rightarrow$ CIN_c + NH_4_c	*pal*
TAL	L-Tyr_c $\rightarrow$ 4-CUA_c + NH_4_c	*tal*
C4H	CIN_c + H^+_c + NADPH_c + O_2_c $\rightarrow$ 4-CUA_c + H_2O_c + NADP_c	*c4h* and *YHR042W*
CL	4-CUA_c + ATP_c + CoA_c + 4.0 H^+_c $\rightarrow$ 4-CUACoA_c + AMP_c + PPi_c	*4cl*
CHS	4-CUACoA_c + 3.0 MaCoA_c $\rightarrow$ NAR-C_c + 3.0 CO_2_c + 4.0 CoA_c + H^+_c	*chs*
CHI	NAR-C_c $\rightarrow$ NAR_c	*chi*
NARt	NAR_c $\leftrightarrow$ NAR_e	-
EX_NAR	NAR_e $\rightarrow$	-

Results

Comparative Flux Sampling Analysis (CFSA)

CFSA was implemented in Python 3.7 using the cobrapy toolbox (v. 0.26.2) and is available in GitLab. Statistical analysis methods based on the Kolmogorov-Smirnov test were implemented using the Scipy stats package. The CFSA output consists of a first Excel file with sampling results for all model reactions that is used as input for filtering. After filtering based on user-defined parameters, a new Excel file containing the filtered results is generated. This file contains the suggested reaction targets and their associated genes as well as possible off-targets caused by multifunctional enzymes. It provides a summary of the sampling results including the mean fluxes in the growth and production scenario as well as the absolute flux change and the reaction equation. The complete files for both case studies are available in GitLab. Additionally, distribution graphs for suggested targets are generated and can be checked before experimental implementation.

Distribution graphs

A distribution graph can be generated for each reaction in the model. It shows the distribution of sampled fluxes in each scenario: growth, production, and slow growth (*i.e.* in how many samples a reaction had a specific flux). Genes are classified as over-expression targets when the absolute mean flux through the corresponding reaction is higher during production than during growth and the flux distributions do not overlap (Figure 3.1 A, B). Similarly, a gene is classified as a down-regulation target when the absolute flux through the corresponding reaction is lower in the production scenario than in the growth scenario and distributions do not overlap (Figure 3.1C, D). The most extreme case of a down-regulation, a knock-out, is obtained when, for a down-regulation target, the flux during production is zero and the gene is not classified as essential (Figure 3.1E, F).

The slow growth scenario is used to reduce the number of incorrectly predicted down-regulation targets (false positives). Growth and production are competing objectives and often low fluxes obtained in the production scenario are not related to increased production but to a decreased growth rate (*e.g.* fluxes through reactions for biomass components). To avoid the identification of these genes as down-regulation targets, reactions for which the production and slow growth distributions overlap are considered false positives (Fig 3.1G).

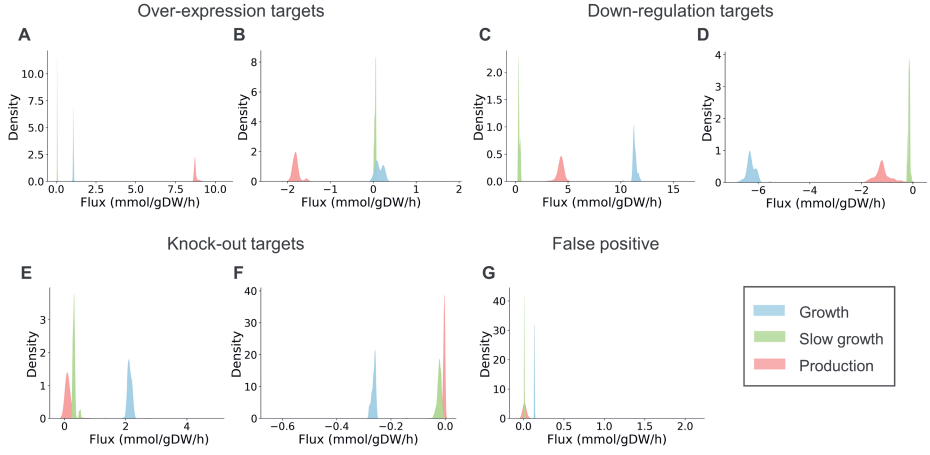


Figure 3.1: Example of distribution graphs illustrating behaviors of candidate targets for metabolic engineering. In each panel, the x-axis represents possible flux values, and the y-axis represents the frequency of each flux value obtained when sampling the solution space, normalized to an area of one. Note that over-expression and down-regulation targets are obtained based on absolute flux values.

Distribution graphs also show the allowed variability of a reaction flux. Reactions with sharp distributions require a specific flux to obtain high production. Conversely, reactions with broad distributions are allowed to carry different fluxes without impacting production and are hence not suitable metabolic engineering targets.

Although targets are automatically filtered based on the overlap between flux distributions in different scenarios, the mean flux, and the range of the distribution during production, visual examination of distribution graphs is recommended as the final filter before experimental implementation.

Effect of sampling parameters on target identification

Samples are taken in three scenarios: growth, slow growth, and production. The user might choose an optimality constraint that determines the minimum growth or production in the growth and production scenarios respectively. The optimality parameter can take values from 0 to 1, where 0 indicates that flux distributions resulting in zero growth or production are allowed, and 1 indicates that only flux distributions with maximum growth or production are allowed. Increasing the optimality parameter tightens the flux constraints, reducing the feasible solution space and decreasing the number of over-expression targets (Figure 3.2). When the optimality parameter is increased, the minimum allowed production increases which results in an increased number of down-regulation targets (Figure 3.2). Reducing the solution space with the optimality parameter ensures that only relevant phenotypes are captured and a value of 0.9 is recommended as default. This optimality value ensures high production while allowing the sampling of sub-optimal phenotypes which increases the robustness of the predictions.

Similarly, a flux faction parameter is used to limit the total sum of fluxes in the production scenario based on the total sum of fluxes in the growth scenario. This parameter can take values equal to or bigger than 1, so higher flux fractions increase the available total flux and, therefore, enlarge the solution space. Although the fraction of unused proteome available for the expression of heterologous pathways is limited and influences production, estimating this fraction is difficult and depends on the growth conditions [169, 170]. Reducing the solution space with the flux fraction parameter as a proxy for proteome constraints reduces the risk of identifying unrealistic loops or excessively long pathways as engineering targets. We show that increasing the flux fraction parameter widens the solution space, increasing the number of down-regulation targets. At the same time, increasing the solution space might result in broader distributions that reduce mean flux differences between growth and production scenarios, reducing the number of possible over-expressions (Figure 3.2). We tested flux fraction values in the 1 to 1.5 range and found 1.25 to be a suitable value for the presented case studies that can be adapted for other applications.

Changing these parameters did not affect the sampling run time which is determined by the model size (94 ± 1 min for the *S. cerevisiae* model and 46 ± 1 min for *C. oleaginosus* model on an Intel(R) Xeon(R) CPU E5-2650 v4 @2.2GHz).

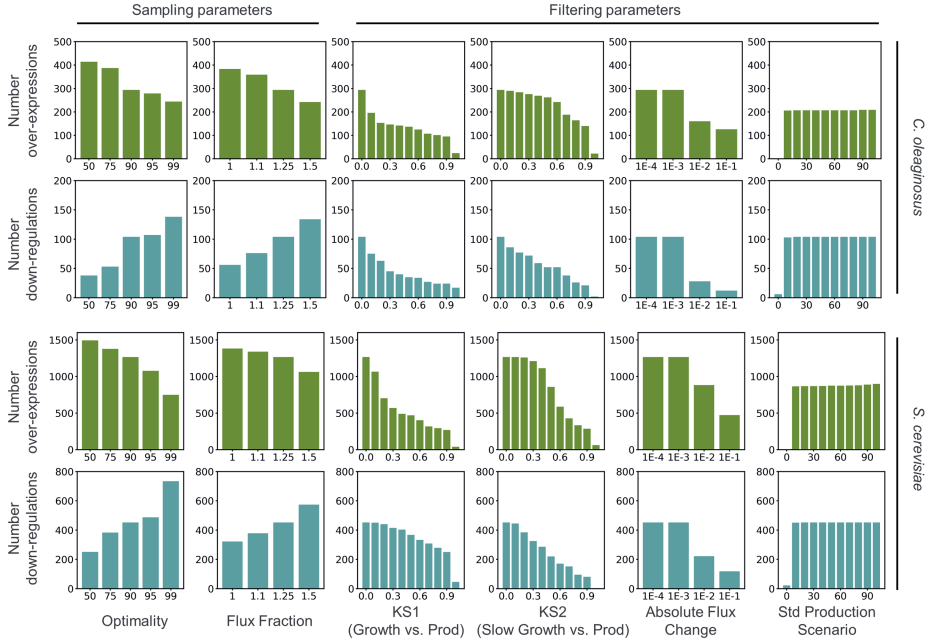


Figure 3.2: Effect of sampling and filtering parameters on the number of reported targets. The effect of sampling parameters was assessed using loose values for the filtering parameters (KS1 = KS2 $\leq$ 1, absolute flux change $\geq$ 0, std production $\leq$ 1000). The effect of the filtering parameters was assessed on samples taken using the recommended sampling parameters (optimality = 90% and flux fraction = 1.25). Filtering parameters not under study were set to KS1 = KS2 $\leq$ 1, absolute flux change $\geq$ 0, std production $\leq$ 1000.

Effect of filtering parameters on target identification

Samples can be evaluated based on the KS test, the mean flux in the different conditions, and the variability of the mean flux in the production conditions (its standard deviation). Here, the effect of these filtering parameters was evaluated using the recommended sampling parameters (optimality = 90% and flux fraction = 1.25).

The KS test identifies whether samples belong to the same distributions (p-value) and the overlap between distributions (KS statistic). Due to the large number of samples, the test is overpowered and p-values do not constitute a good filtering criteria even if corrections for multiple testing such as Bonferroni are applied. The KS statistic determines the overlap between distributions, where a KS value of zero indicates complete overlap. This parameter largely affects the number of selected targets. KS1 indicates the overlap between the growth and production scenarios and ensures significantly different flux distributions among these conditions. Therefore, increasing KS1 reduces the number of over-expression and down-regulation targets (Figure 3.2). KS2 indicates the overlap between the slow growth and production scenarios. This condition is used to distinguish reactions in which flux decreases due to the decreased growth in the production scenario instead of the

increased production. Therefore, increasing KS2 reduces the number of down-regulation targets. As default, the use of 0.75 for KS1 and KS2 is recommended but this value can be tuned based on the distribution of KS values obtained after sampling (see example in GitLab).

The absolute flux change between growth and production scenarios is used to rank the targets and can be used as an additional filter (Figure 3.2). Targeting reactions with considerable flux changes when growth or production are maximized, increases the chance of significant phenotype changes *in vivo*. Contrarily, low flow changes are unlikely to be achieved by adjusting gene expression. A value of 0.01 mmol/g_{DW}/h is used as default. Although we recommend the use of mean absolute change to favor reactions with larger fluxes, CFSA allows alternative filtering based on mean fold changes and mean relative changes of reaction fluxes.

The standard deviation of the flux distributions in the production scenario is used as an additional filter and reactions with broad distributions are not considered relevant targets. Higher standard deviations indicate that production is not affected by the flux of the studied reaction. Decreasing the maximum allowed standard deviation therefore reduces the number of suggested targets for up and down-regulation. Considering that reactions with high mean fluxes also have a higher standard deviation, the default value of this parameter is set to 50.

Case studies

The number of possible metabolic-engineering targets obtained using CFSA for lipid production in *C. oleaginosus* and naringenin production in *S. cerevisiae* is presented in Table 3.2. These values were obtained using optimality = 0.90, flux fraction = 1.25, KS1 = KS2 $\geq$ 0.75, mean absolute change $\geq$ 0.01, and standard deviation in production $\leq$ 50. The complete list of target reactions can be found in GitLab (filtered_results.xlsx). The sections below elaborate on some of the targets found.

Table 3.2: Number of targets obtained in the *C. oleaginosus* and *S. cerevisiae* case studies.

	C. oleaginosus	***S. cerevisiae***
Number over-expressions	25	50
Number down-regulations	1	41

Production of lipids by *C. oleaginosus*

C. oleaginosus is an oleaginous yeast able to accumulate lipids above 40% (w/w) of its biomass when growing under nitrogen limitation [81]. The composition of the produced fatty acids is comparable to commonly used plant-derived oils. Therefore it has been flagged as an auspicious microbial cell factory for sustainable lipid production at an industrial scale [171, 172]. The lipid accumulation initiates with the transport of citrate from the mitochondria to the cytosol where it is cleaved into acetyl-CoA by ATP citrate lyase (ACL). Acetyl-CoA is converted into malonyl-CoA by acetyl-CoA carboxylase (ACC) leading to the activation of the lipid synthesis and elongation pathways (Fig 3.3 A) [173, 174].

CFSA was applied to investigate metabolic engineering strategies for enhanced lipid production in *C. oleaginosus*. As a result, we obtained one candidate reaction for down-regulation and 25 candidate reactions (belonging to 11 groups) for over-expression. The complete list of reactions is available in GitLab (filtered_results.xlsx). The only down-regulation target found corresponds to ATP diphosphohydrolase which catalyzes the conversion of ATP to AMP and reflects the high energy requirements of lipid production compared to growth.

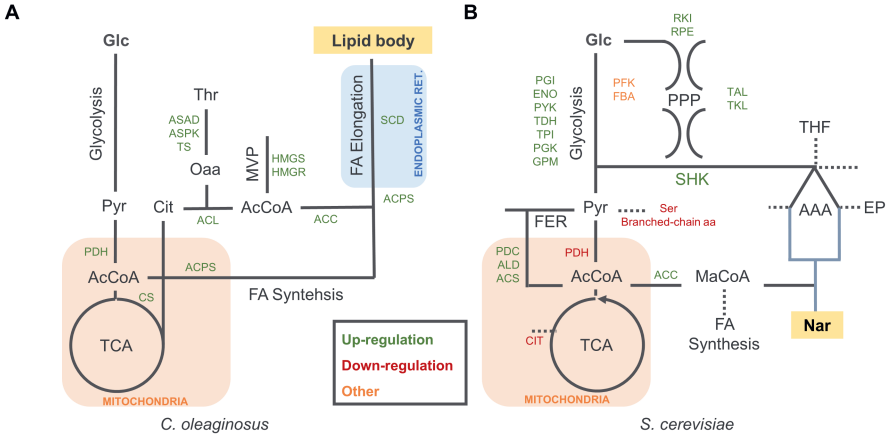


Figure 3.3: Summary of selected metabolic engineering targets for lipid production in *C. oleaginosus* (**A**) and naringenin (NAR) production in *S. cerevisiae* (**B**). Endogenous and heterologous metabolic pathways are simplified and presented in grey and blue respectively: TCA, tricarboxylic acid cycle; MVP, mevalonate pathway; FA, fatty acids; PPP, pentose phosphate pathway; SHK, shikimate pathway; FER, fermentative pathway; AAA, aromatic amino acid pathway; EP, Ehrlich pathway. *Metabolite abbreviations*: Glc, glucose; Pyr, pyruvate; AcCoA, acetyl-coenzyme A; Cit, Citrate; MaCoA, malonyl-coenzyme A. *Reaction abbreviations*: PDH, pyruvate dehydrogenase; CS, citrate synthase; ASAD, aspartate-semialdehyde dehydrogenase; ASPK, aspartate kinase; TS, threonine synthase; ACL, ATP citrate lyase; ACPS, fatty acyl-ACP synthase; HMGS, hydroxymethylglutaryl-CoA synthase; HMGR, hydroxymethylglutaryl-CoA reductase; ACC, acetyl-CoA carboxylase; SCD, stearoyl-CoA desaturase; PGI, glucose-6-phosphate isomerase; ENO, enolase; PYK, pyruvate kinase; TDH, glyceraldehyde-3-phosphate dehydrogenase; TPI, triose phosphate isomerase; PGK, phosphoglycerate kinase; GPM, 2,3-bisphosphoglycerate-independent phosphoglycerate mutase; PDC, pyruvate decarboxylase; ALD, aldehyde dehydrogenase; ACS, acetate-CoA ligase; CIT, citrate synthase; PFK, phosphofructokinase; FBA, fructose-6-phosphate aldolase; RKI, ribose-5-phosphate isomerase; RPE, ribulose-phosphate 3-epimerase; TAL, transaldolase; TKL, transketolase.

As expected, reactions from the fatty acid synthesis pathway (fatty acyl-ACP synthase (ACPS) and stearoyl-CoA desaturase (SCD)) were suggested as over-expression targets (Figure 3.3A). Besides, key reactions ACL and ACC, which over-expression has improved lipid accumulation in other oleaginous yeast, were also predicted as targets [175, 176]. Additionally, pyruvate dehydrogenase (PDH), which connects the glycolytic pathway to the TCA cycle, and citrate synthase (CS),

which synthesizes citrate from acetyl-CoA and oxaloacetate, were predicted as up-regulation targets to improve cytoplasmic citrate supply for lipid synthesis. However, CFSA did not predict down-regulation targets from the β-oxidation pathway that are commonly used to increase the availability of acetyl-CoA for fatty acid synthesis [177, 178].

Along with these experimentally validated targets, unanticipated reactions from the mevalonate pathway such as hydroxymethylglutaryl-CoA synthase (HMGS) were suggested as beneficial up-regulations by CFSA. Although these reactions consume acetyl-CoA, simulations that reduce HMGS flux result in diminished lipid production (Sup. Figure 3.1). The complex relationship between the mevalonate pathway and lipid synthesis, intrinsic to the GEM used, is captured by CFSA as a possible strategy to increase fatty acid production.

Finally, CFSA suggested over-expression targets from amino acid metabolism, including reactions involved in threonine synthesis (aspartate-semialdehyde dehydrogenase (ASAD), aspartate kinase (ASPK), threonine synthase (TS)), as well as glutamate and glutamine synthesis. Threonine synthesis requires cytosolic oxaloacetate which is produced during citrate conversion to acetyl-CoA by ACL. Therefore, the over-expression of the threonine synthesis pathway could improve lipid production balancing the over-production of oxaloacetate. Furthermore, Kim et al. suggested that upregulation of threonine synthesis could potentially increase the fluxes through the TCA cycle which increases citrate supply for acetyl-CoA synthesis [179].

Production of naringenin by *S. cerevisiae*

S. cerevisiae is a model organism with in-depth genetic and physiological characterization, ample application in industrial bioprocesses, and Generally Regarded as Safe (GRAS) status [180, 181]. Flavonoids such as naringenin are precursors of anthocyanins and have traditionally been used for fragrance, flavor, and color in various food types [182, 183]. Naringenin is derived from the shikimate pathway for aromatic amino acid biosynthesis, which starts with the condensation of erythrose-4-phosphate (E4P) and phosphoenolpyruvate (PEP) (Figure 3.3 B). Production of this compound requires the heterologous expression of 4CL, CHS, CHI, and either PAL, C4H, and CPR for production from phenylalanine, or TAL for production from tyrosine (Table 3.1). Moreover, naringein production requires an appropriate supply of malonyl-CoA [184].

We used CFSA to find metabolic engineering strategies that improve naringenin production. We obtained 41 reaction candidates for down-regulation (belonging to 28 groups) and 50 targets for up-regulation (belonging to 35 groups). The complete list of reactions is available in GitLab (filtered_results.xlsx). Two proteomic datasets covering 48.4% of the genes and 23.6% of the reactions in the model were used as an additional filtering step to prioritize the 34 detected proteins as down-regulation targets. Out of the 50 up-regulation targets, 33 targets can be considered obvious as they are part of the production pathway for naringenin or its precursors and/or have been experimentally tested. Two of the down-regulation targets, prioritized by the proteomic datasets, have been tested *in vivo*.

As expected, reactions from the shikimate and naringenin production pathways were predicted as over-expression targets, with a preference for the tyrosine branch. Reactions belonging to glycolysis and non-oxidative pentose phosphate pathway were also predicted as suitable targets (Figure 3.3B). Notably, CFSA suggested a decreased flux through the phosphofructokinase (PFK) and aldolase (FBA) reactions that result in the conversion of fructose-6-phosphate (F6P) to fructose-1,6-biphosphate (F1,6bP) and, subsequently to dihydroxyacetone-phosphate (DAP) and glyceraldehyde-3-phosphate (GAP). Instead, it proposed F6P conversion to E4P and DAP via sedoheptulose-1,7-biphosphate (S1,7bP). In this way E4P and DAP, which can be later converted to PEP, are simultaneously produced, feeding the shikimate pathway with its two precursors. In the cells, PFK and FBA are responsible for both reactions and conversion of F6P to DAP, and GAP is favored due to a higher affinity of FBA towards F1,6-bP. According to the simulations, expressing *pfk* genes with higher affinity towards S7P such as *ppi-pfk* from *Clostridium thermosuccionogenes* [185] is suggested as a novel strategy to improve naringenin production.

CFSA also suggested experimentally validated strategies to increase the production of acetyl-CoA and malonyl-CoA syntheses such as the down-regulation of PDH and CIT2 and the up-regulation of PDCS, ALD, ACS, and ACC (Fig 3.3B) [186]. However, it fails to predict the down-regulation of fatty acid synthesis as a strategy to improve malonyl-CoA availability [181, 186, 187]. Similarly, CFSA does not suggest the deletion of Ehrlich pathway (EP) genes, involved in the degradation of intermediates that have been shown to improve the production of other aromatic compounds [183, 188, 189, 190].

The model suggests down-regulation of reactions involved in serine and branched-chain amino acid synthesis, likely to decrease the conversion of pyruvate into biomass components. Besides, the down-regulation of reactions involving tetrahydrofolic acid (THF) and glycine is also suggested, probably to reduce chorismate consumption for THF formation (Fig 3.3B).

Last, CFSA suggested the over-expression of adenylate kinase (ADK) which has been reported to increase the production of malonyl-CoA-derived products [191]. Similarly, although experimentally unrealistic, CFSA suggested down-regulation of ATP synthase and reactions from the electron transport chain, reflecting the lower energy requirements of production compared to growth.

Discussion

Inspired by other strain design algorithms that suggest metabolic engineering targets based on the comparison of predicted fluxes between wild type and production phenotypes [152, 153, 154, 155], we present CFSA. This tool is based on flux sampling and, by analyzing the complete GEM solution space, can guide the design of microbial cell factories. While current tools focus on design approaches for growth-coupled production, we present a first tool that allows the design of growth-uncoupled production strategies. Besides, as opposed to previous tools, CFSA designs are based on the complete exploration of the solution space achieved using flux sampling instead of non-unique FBA solutions, which ensures the full inspection of cell metabolism [64].

We applied CFSA to improve the production of lipids by *C. oleaginosus*, an endogenous product in an emerging cell factory, as well as the production of naringenin in *S. cerevisiae*, a heterologous product in an established industrial microorganism. In both cases, CFSA suggested experimentally validated targets including evident up-regulations belonging to the product synthesis pathway and distant targets that improve precursor availability. It also suggested new engineering strategies involving alternative pentose phosphate pathway reactions for naringenin synthesis and the over-expression of genes from the threonine and mevalonate pathways for lipid production. Although the lack of mechanistic understanding of some of the suggested targets could question their implementation, it also highlights the potential of model-driven approaches to find non-obvious engineering strategies.

Regarding down-regulations, CFSA failed to predict reported successful targets from pathways involved in the degradation of production pathway intermediates (*e.g.* Ehrlich pathway genes deletion for naringenin production [183, 188, 189, 190] or β-oxidation gene knock-outs for improved lipid production [177, 178]). This pitfall is shared with other GEM-based strain design approaches since, although active *in vivo*, these reactions remain inactive in GEM simulations.

In addition to the simulation of growth and production phenotypes, we include the simulation of slow growth. This scenario is used to differentiate between fluxes that change due to the simulated low growth in the production scenario, from fluxes that are potentially related to increased production (Figure 3.1). The simulation of this phenotype is used as negative control reducing false positive target predictions (Figure 3.2). Other tools use FBA to maximize growth or production and use the obtained solution to compare reaction fluxes and identify engineering targets. Instead, we include the optimality parameter to ensure that not only the optimal flux profile is sampled, but less efficient behavior is also tolerated potentially leading to increased robustness.

CFSA is easy to implement and the filtering criteria used results in a reduced list of potential targets for up-, down-regulation, and knock-outs for subsequent inspection. Default parameters can easily be modified according to user needs and additional filtering criteria such as the use of proteomic data can be integrated into the workflow. Other strain design methods provide minimal intervention strategies that, although useful in theory, might not return the expected results in practice. For example, when entire pathways are predicted as up-regulation targets, GEM cannot identify limiting reactions as enzyme kinetics and regulation are not considered in constraint-based models [149]. Instead, we provide a complete list of possible interventions and endow the user with additional information to make a decision based on the most feasible suggestions.

Disadvantages of CFSA include the difficulty of estimating the quantitative effect of the suggested manipulations, which are identified based on statistical testing. There is no guarantee that the effect size in flux correlates to the strength of the knock-down or over-expression required. Besides, targets are suggested as individual interventions and the effect of possible combinations of targets is not provided. Still, after a first round of CFSA, the algorithm can be re-applied using GEMs with modified reaction bound that simulate the desired interventions to find potential complementary targets.

As with other methods based on GEMs, errors in the models can lead to unexpected outcomes. For example, loops in the metabolic network can lead to artificially high fluxes and thus false predictions. Often parsimonious flux balance analysis [161] or loop-less flux balance analysis [192, 193] are used to limit these errors, however, implementing these in flux sampling is non-trivial. We mimicked the parsimonious approach by adding a flux fraction parameter that constrains the total sum of fluxes in the production scenario based on the growing phenotype. However, this approach, while biologically reasonable for the reference condition at a high growth rate, does not necessarily apply to the production condition, as the assumption of minimal flux and thus protein usage does not hold for this artificial scenario. Instead, loop-less flux sampling approaches such as the loop-less Artificially Centered Hit-and-Run on a Box algorithm (ll-ACHRB) [194] or the LooplessFluxSampler [195], currently implemented in Matlab, could be used as alternatives.

CFSA is the first strain design algorithm based on flux sampling that explores the whole solution space of a GEM and suggests metabolic engineering targets for growth-uncoupled production. Its robustness, simplicity, and flexibility make it ideal to complement and systematize the design of microbial cell factories. CFSA predictions, including non-obvious targets, can be sequentially tested using high-throughput approaches such as automated platforms and biosensor-aid screening accelerating and broadening the strain design process.

Declaration of interest

The authors declare no conflict of interest.

Acknowledgment

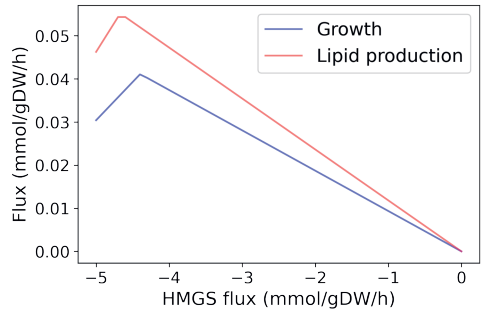
This project has been funded by the Netherlands Organization for Scientific Research (NWO; project number GSGT.2019.008) and the Dutch Ministry of Agriculture through the "TKI-toeslag" project LWV19221 "Tailor-made microbial oils and fatty acids".

Data availability

Scripts and data are available at Gitlab.

GitLab

Supplementary Figures



Sup. Figure 3.1: Effect of HMGS flux on growth and production fluxes in simulations with the *i*NP636_*Coleaginosus*_ATCC20509 GEM. Fluxes were obtained using flux balance analysis fixing the bounds of the HMGS reaction (r_0599) and setting the growth (Biomass_nitrogen_depletion) or lipid production (Ex_lipid_body_cytosol) reactions as objective to maximize.

CHAPTER 4

Shikimate pathway-dependent catabolism enables high-yield production of aromatics

Lyon Bruinsma*, Christos Batianis*, Sara Moreno Paz, Kesi Kurnis, Job J. Dirkmaat, Ruud A. Weusthuis, Vitor A. P. Martins dos Santos

*Contributed equally

This chapter is ready for submission

Abstract

The biotechnological application of microorganisms to replace fossil-based processes is a global necessity. To fully leverage their potential, high titers, rates, and yields are required for commercial processes. This generally involves radical reprogramming of the intrinsic metabolism. In this study, we demonstrate a combination of metabolic modeling, rational engineering, and adaptive laboratory evolution to radically refactor bacterial metabolism. We created a new-to-nature shikimate pathway-dependent catabolism in *Pseudomonas putida* by reprogramming the shikimate pathway as the dominant pathway for growth. This new strain diverts the vast majority of its carbon catabolism flux through the shikimate pathway and produces 0.35 mol/mol 4-hydroxybenzoate in glycerol minimal medium during growth, achieving 89.2% of the maximum predicted pathway yield. We demonstrate that the shikimate pathway can act as the main catabolic route and deliver a promising strain that can serve as a useful *chassis* to produce various shikimate pathway-derived compounds.

Introduction

Metabolism defines the lifestyle of any organism [196]. Despite the considerable diversity among organisms, almost all of them share the primary central carbon metabolism. This primary metabolism, often consisting of glycolysis, the tricarboxylic acid (TCA) cycle, and the pentose phosphate pathway is used to convert carbon sources to precursor molecules necessary to synthesize all cellular constituents for growth and maintenance [197]. As a result, the primary metabolism is often viewed as a rigid network, and any deviation from it can lead to reduced cell viability or even cell death [198]. This is unfortunate as we can leverage this network to produce a plethora of fuels and chemicals, replacing petroleum-derived processes [199]. Therefore, for the sake of the bioeconomy, we need to be able to efficiently manipulate these networks to produce chemicals with high titers, rates, and productivities [200]. To achieve this goal, microorganisms need to be radically refactored to ensure a high flux towards the product of interest. Yet, the introduction of a production pathway often interferes directly with the main metabolism, and fluxes are not easily diverted [201]. Moreover, extra complexity is added as many inherent pathways are subjected to tight regulation and therefore carry considerably low fluxes [200]. One example is the shikimate pathway, the biochemical source of numerous aromatic molecules including aromatic amino acids. While a myriad of valuable molecules can be derived from this pathway, its biotechnological exploitation remains a mounting metabolic engineering challenge [202, 203, 204].

Standard metabolic engineering strategies rely on simple gene overexpressions and flux alterations in the central carbon metabolism. However, these modifications compete directly with cellular fitness and often do not result in economically feasible yields [205]. Therefore, a paradigm shift is essential to fully reshape the rigid carbon metabolism and exert the potential of the shikimate pathway. In recent years, growth-coupled selections have emerged as powerful tools to redesign cell factories for the production or consumption of new substrates or to establish new metabolic architectures [206, 207, 208, 209]. This approach combined with laboratory evolution has the power to completely reprogram cellular metabolism, creating industrially relevant cell factories. One key example is the establishment of a chemo-autotrophic *Escherichia coli* that can generate all its biomass from CO_2 [210]. By introducing essential deletions in xylose catabolism, bacterial growth became dependent on the carboxylation reaction by Rubisco. Eventually, this dependency was able to establish full autotrophic growth after several rounds of laboratory evolution. Using a similar approach, a synthetic methylotrophic *E. coli* strain was engineered [211]. Here, methanol utilization was coupled to xylose catabolism, which after evolution yielded a strain that could generate all biomass and energy from this promising C1-feedstock. Apart from introducing foreign pathways in a microbial host, growth-coupled selection systems can be used to fully rearrange native metabolic architectures. Iacometti et al. demonstrated the flexibility of bacterial metabolism by establishing silent glycolytic routes in *E. coli* [212]. Using growth-coupled evolution they acquired a strain in which the canonical Embden-Meyerhof-Parnas (EMP) pathway was replaced by a serine shunt. Similarly, these growth-coupled scenarios can be deployed as a metabolic engineering strategy for the shikimate pathway.

Here, we describe how we rigorously rearranged the metabolic network of the industrially relevant bacterium *Pseudomonas putida*, creating a shikimate pathway-dependent catabolism (SDC) (Figure 4.1). We established a model-driven rearrangement of the main carbon metabolism through pyruvate-driven laboratory evolution using innate pyruvate-releasing reactions from the shikimate pathway with the aid of a selective biosensor. Whole genome sequencing and reverse engineering revealed that a perturbation in the signaling network was key in realizing this drastic metabolic shift. Further optimization of SDC was achieved using rational and model-driven approaches, resulting in the first strain ever constructed which uses the shikimate pathway as the dominant catabolic pathway, serving as a useful *chassis* to produce various aromatic compounds. Our findings highlight the tremendous plasticity of metabolic networks and how growth-coupled strategies can be exploited to install new-to-nature metabolisms for industrial biotechnology.

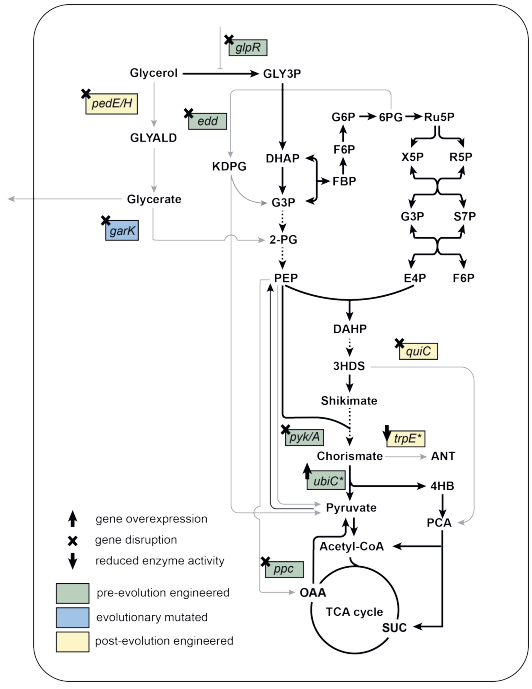


Figure 4.1: Metabolic architecture of the shikimate pathway-dependent catabolism (SDC). Pathways and mutations that were engineered pre- and post-evolution to establish the final SDC strain. Abbreviations: GLY3P, glycerol-3-phosphate; GLYALD, glyceraldehyde; DHAP, dihydroxyacetone phosphate; G3P, glyceraldehyde-3-phosphate; 2-PG, 2-phosphoglycerate; KDPG, 2-keto-3-deoxy-6-phosphogluconate; PEP, phosphoenolpyruvate; FBP, fructose bisphosphate, F6P, fructose-6-phosphate; G6P, glucose-6-phosphate; 6PG, 6-phosphogluconate; Ru5P, ribulose-5-phosphate; R5P, ribose-5-phosphate; X5P, xylose-5-phosphate; S7P, sedoheptulose-7-phosphate; E4P, erythrose-4-phosphate; DAHP, 3-deoxy-d-arabinoheptulosonate-7-phosphate; 3HDS, 3-dehydroshikimate; ANT, anthranilate; 4HB, 4-hydroxybenzoate; PCA, protocatechuate; SUC, succinate; OAA, oxaloacetate; *glpR*, glycerol regulon repressor; *edd*, phosphogluconate dehydratase; *pedE/H*, PQQ - dependent alcohol dehydrogenases; *garK*, glycerate kinase; *pykA*, pyruvate kinase; *quiC*, dehydroshikimate dehydratase; *ubiC*, chorismate pyruvate lyase; *ppc*, phosphoenolpyruvate carboxylase; *trpE*, anthranilate synthase.

Material and methods

Plasmids, primers, and strains

All strains and plasmids used in the present study are listed in Sup. Table 1. Primers used for plasmid construction and gene deletions are listed in Sup. Table 2.

Culture conditions and medium

P. putida and *E. coli* cultures were incubated at 30°C and 37°C respectively. For cloning purposes, both strains were propagated in Lysogeny Broth (LB) medium containing 10 g/l NaCl, 10 g/l tryptone, and 5 g/l yeast extract. For the preparation of solid media, 1.5% (w/v) agar was added. Antibiotics, when required, were used at the following concentrations: kanamycin (Km) 50 µg/ml, gentamycin (Gm) 10 µg/ml, chloramphenicol (Cm) 50 µg/ml and apramycin (Apra) 50 µg/ml. All growth experiments were performed using M9 minimal medium (per liter; 3.88 g K_2HPO_4, 1.63 g NaH_2PO_4, 2.0 g $(NH_4)_2SO_4$, pH 7.0). The M9 media was supplemented with a trace elements solution (per liter; 10 mg ethylenediaminetetraacetic acid (EDTA), 0.1 g $MgCl_2{\cdot}6H_2O$, 2 mg $ZnSO_4{\cdot}7H_2O$, 1 mg $CaCl_2{\cdot}2H_2O$, 5 mg $FeSO_4{\cdot}7H_2O$, 0.2 mg $Na_2MoO_4{\cdot}2H_2O$, 0.2 mg $CuSO_4{\cdot}5H_2O$, 0.4 mg $CoCl_2{\cdot}6H_2O$, 1 mg $MnCl_2{\cdot}2H_2O$). Strains were precultured in 10 ml LB with corresponding antibiotics. Then, the cultures were washed twice in M9 media without a carbon source. Finally, the cultures were diluted to an OD600 of 0.1 to start the experiment. Flask experiments were performed in 250 ml Erlenmeyer flasks filled with 25 ml of M9 minimal medium with 40 (SDC characterization) or 200 (ALE experiment) mM glycerol and the cultures were incubated in a rotary shaker at 200 rpm at 30°C.

Cloning procedures

Plasmids were constructed using the standard protocols of the previously described SevaBrick assembly [126]. All DNA fragments were amplified using Q5® Hot Start High-Fidelity DNA Polymerase (New England Biolabs). To construct the plasmids pSensor and pSensorEvo, the 4-hydroxybenzoate (4HB)-responsive regulator and promoter (PobR/PpobA) were synthesized by Integrated DNA Technologies (IDT), and the *ubiC*$^{E31Q/M34}$ was amplified from the genome of *E. coli* with primers to introduce the E31Q and M34V mutations. All parts were integrated into the pSB1C3 repository and subsequently assembled into pSEVAb83 following the standard procedures of the SevaBrick assembly. All plasmids were transformed via heat shock in chemically competent *E. coli* DH5α λ*pir* cells and via electroporation or conjugation in *P. putida* (Thermo Fisher Scientific). Transformants were selected on LB agar plates with corresponding antibiotics and colonies were tested by colony PCR with Phire Hot Start II DNA polymerase (Thermo Fisher Scientific). After extraction, all constructs were verified by Sanger DNA sequencing (MACROGEN Inc.).

Genome modifications

Genomic modifications were performed as described by Wirth et al. [213]. Homology regions of ± 500 bp were amplified up and downstream of the target gene from the genome of *P. putida* KT2440. Both regions were cloned into the non-replicative pGNW vector and propagated in *E. coli* DH5α λ*pir*. Correct plasmids were transformed into *P. putida* by electroporation and selected on LB + Km plates. Successful co-integrations were verified by PCR. Hereafter, co-integrated strains were transformed with the pQURE6-H, and transformants were plated on LB + Gm containing 2 mM 3-methylbenzoic acid (3-mBz). This compound induces the XylS–dependent Pm promoter, regulating the I-SceI homing nuclease that cuts the integrated pGNW vector. Successful gene deletions were verified by PCR and Sanger sequencing (MACROGEN inc). The P290S modification in *trpE* was verified by MASC PCR as described by Wang & Church [214]. Hereafter, the pQURE6-H was cured by removing the selective pressure and its loss was verified by sensitivity to gentamycin.

Analytical methods

Cell growth was determined by measuring the optical density at 600 nm (OD600) using an OD600 DiluPhotometer spectrophotometer (IMPLEN) or a Synergy plate reader (BioTek Instruments). Analysis of glycerol and pyruvate in supernatants was performed using high-performance liquid chromatography (HPLC) (Thermo Fisher Scientific) equipped with an Aminex HPX-87H column. The mobile phase was 5 mM of H_2SO4_4 at a flow rate of 0.6 ml/min, the column temperatures were held at 60 °C and the compounds were detected using a Shodex RI-101 detector (Shodex). The amount of produced 4HB was determined using HPLC (Shimadzu) with a C18 column (4.6 mm × 250 mm) and a UV/vis detector set at 472 nm. The mobile phase consisted of Milli-Q water (A), 100 mM formic acid (B), and acetonitrile (C) with a flow rate of 1 ml/min at 30°C. Chromatographic separation of analytes was attained using the following gradient program: t = 0 - 5min: A-55%, B-10%, and C-35%; from t = 5 - 10min ramp to A-10%, B-10%, and C-80% and held until 15 min. Then from t = 15 – 16 min, the gradient was returned to A-55%, B-10%, and C-35% and maintained isocratic for a total run time of 18 min. For quantification, calibration curves were prepared using pure standards (99% purity) purchased from Sigma-Aldrich.

Adaptive laboratory evolution

Strain Δ*pyr* was inoculated in two 250 ml shake flasks with M9 minimal medium with 200 mM glycerol as the sole carbon source and incubated in a rotary shaker at 200 rpm at 30°C. The starting cell density was set at OD600 = 0.1 and cells were diluted back to the same OD600 after they reached an OD600 of >1.0. Evolved strains were selected on M9 agar plates with 40 mM glycerol and characterized in 200 µl of M9 medium with 40 mM glycerol using a Synergy plate reader (BioTek Instruments). Cell density (OD600) and GFP fluorescence (excitation 485 nm, emission 512 nm, gain 50) were measured over time using continuous linear shaking (567 cpm, 3mm). Growth rates were calculated by taking the natural log of the OD600 values.

Whole-genome sequencing

The genomic DNA of the evolved mutants was isolated from LB overnight cultures using the GenElute™ Bacterial Genomic DNA Kit (Sigma-Aldrich St. Louis, MO). The extracted DNA was evaluated by gel electrophoresis and quantified by a NanoDrop spectrophotometer (Thermo Fisher Scientific). Samples were sent for Illumina sequencing to Novogene Co. Ltd. (Beijing, China). Raw Illumina reads were trimmed for low quality and adapters with fastp (v0.20.0) [215]. Mutations were identified by comparing the reads to the annotated reference genome of *P. putida* KT2440 (GCF_000007565.2) using breseq (v0.35.5) [216].

Genome-scale metabolic modeling

Computational analysis was performed using COBRApy (v0.18.1) and Python (v3.6). We used iJN1462, the latest developed genome-scale model (GEM) of *P. putida* to rank pyruvate-releasing reactions and to simulate native and SDC metabolism [217]. In all simulations, glycerol was used as the sole carbon source with a maximum uptake rate of 3.95 mmol/g_{DW}/h [218]. All metabolic reactions able to produce pyruvate in the iJN1462 GEM were evaluated for their capacity to support growth as the sole pyruvate source. The upper and lower bounds of all pyruvate-releasing reactions were constrained to zero except for two essential reactions: ANS2 (*trpE*, *pabA*) and ADCL (*pabC*). Iteratively, the flux through each reaction was unconstrained making it the main available source of pyruvate in the model. The growth rate, represented by the BIOMASS_KT2440_WT3 reaction, was maximized. Reactions were ranked according to the predicted maximum growth rates relative to the wild-type growth rate (100%). iJN1462 was modified to correctly simulate SDC and wild-type metabolism according to Batianis et al. [219]. Briefly, reactions AKGDb, THRA, LSERDHr, ACACT1r, MACCOAT, MMSAD3, ACALD, ALDD2x, ALDD2y and KAT1 were made irreversible, and the stoichiometry of the GAPDi_nadp reaction was corrected (see Sup. Table 3 for details). When simulating SDC, a parsimonious FBA-like constraint was applied such that the sum of all the predicted fluxes cannot exceed the sum of all predicted fluxes in native metabolism. Besides, as a base for SDC simulation, we constrained to zero the flux through reactions EDD, PYK, AGPOP, ME2, OAAFC, SERD_L, CYSTL, CYSDS, MCITL2, LDH_2, and LDH_D2 to reproduce the result of the biosensor-assisted ALE that made CHRPL the main pyruvate source. Model modifications that allowed the simulation of the different SDC strains are presented in Sup. Table 4. Maximum theoretical 4HB yields were calculated maximizing the 4HB exchange reaction (EX_4hbz_e). The optimal metabolism of the SDC strains and their maximum growth rates were simulated maximizing biomass production (BIOMASS_KT2440_WT3).

Statistical analysis

All reported experiments are derived from independent biological replicates. Figures represent the mean values of corresponding biological triplicates and the standard deviation. Significant differences among results were evaluated by unpaired Student's t-tests.

Results

Metabolic design and *in silico* assessment of shikimate-dependent catabolism

Pyruvate is a key node in central metabolism predominantly produced in *P. putida* via the Entner-Doudoroff (ED) and lower Embden–Meyerhof–Parnas (EMP) pathways. Yet apart from the main carbon metabolism, pyruvate is produced by several other innate reactions, some derived from the shikimate pathway. Therefore, to increase the flux through the shikimate pathway, we aimed to establish it as the main source of cellular pyruvate creating a shikimate-dependent catabolism (SDC). We used iJN1462, a genome-scale model (GEM) of *P. putida*, to find all the pyruvate-releasing reactions present in its genome. We analyzed the ability of each of these reactions to support *in silico* growth as the sole pyruvate source using glycerol as substrate (Figure 4.2A). The model predicted the highest growth rate using a shikimate pathway reaction when chorismate pyruvate lyase (CHRPL) was the sole pyruvate source. This reaction cleaves chorismate, the final product of the shikimate pathway, to pyruvate and 4-hydroxybenzoate (4HB) (Figure 4.2B). When CHRPL is the sole pyruvate source, the *in silico* growth rate is reduced by 17.8% compared to the wild type (*P. putida* KT2440). Still, it allows 26.1% faster growth than the other pyruvate-releasing shikimate reactions and was therefore chosen as the most efficient candidate to establish SDC.

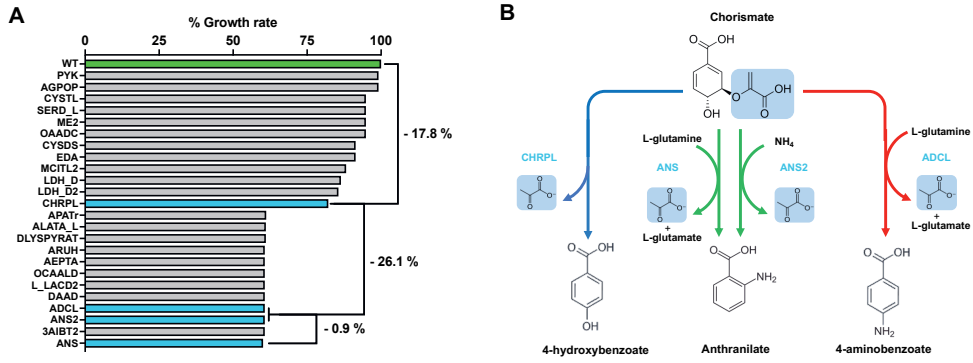


Figure 4.2: *In silico* assessment of SDC. **A.** Growth rate predictions using FBA with glycerol as carbon source when each pyruvate-releasing reaction is set as the sole pyruvate source. Growth rates relative to the wild type growth rate (100%, μ_{max}= 0.235 h^{-1}) are shown. **B.** Metabolic scheme of the pyruvate-releasing reactions derived from the shikimate pathway. PYK, pyruvate kinase; AGPOP, DGTP:pyruvate 2-O-phosphotransferase; CYSTL, cystathionine b-lyase; SERD_L, L-serine deaminase; ME2, malic enzyme (NADP); OAADC, oxaloacetate decarboxylase; CYSDS, cysteine desulfhydrase; EDA, 2-dehydro-3-deoxy-phosphogluconate aldolase; MCITL2, methylisocitrate lyase; LDH_D, D-lactate dehydrogenase; LDH_D2, D-lactate dehydrogenase (q8); CHRPL, chorismate pyruvate lyase; APATr, B-alanine pyruvate aminotransferase; ALATA_L, L-alanine transaminase; DLYSPYRAT, D-lysinepyruvate aminotransferase; ARUH, L-arginine pyruvate transaminase; AEPTA, 2-aminoethylphosphonate pyruvate transaminase; OCAALD, 4-oxalcitromalate aldolase; L_LACDS, L-lactate dehydrogenase (ubiquinone); DAAD, D-amino acid dehydrogenase; ADCL, 4-aminobenzoate synthase; ANS2, anthranilate synthase 2; 3AIBT2, L-3 aminoisobutyrate transaminase; ANS, anthranilate synthase.

To further evaluate the feasibility of SDC, we calculated the overall chemical equations from glycerol to pyruvate in terms of its ability to generate reduced cofactors and ATP compared to the native metabolism (see Sup. Methods). For this purpose, native and SDC metabolism were simulated with flux balance analysis (FBA) in which pyruvate production was set as the main objective using glycerol as the carbon source. During native metabolism, glycerol is equimolarly converted to pyruvate, generating reducing equivalents in the process. As *P. putida* is an aerobic bacterium, a large surplus of ATP can be generated in the electron transport chain. In contrast, the SDC metabolism is energetically poor compared to the native metabolism. The shikimate pathway is an anabolic pathway and requires the incorporation of NADPH and ATP to reach the end product chorismate. Moreover, there is a net production of CO_2 reducing the pyruvate yield. Nonetheless, with the production of reducing equivalents, a surplus of 0.23 ATP can be generated for growth and maintenance.

Creating a growth-coupled scenario to establish SDC

Our *in silico* analysis demonstrates the feasibility of SDC. However, prediction relied solely on stoichiometry. Microbial metabolism is tightly regulated, and a metabolic reconfiguration of this extent is impossible to attain through rational engineering. Adaptive laboratory evolution (ALE) is a microbial engineering method commonly used to achieve desired phenotypes that cannot be obtained using the rational approach [220]. Therefore, to install CHRPL as the major pyruvate source, we established a pyruvate auxotrophic strain (Δ*pyr*) to allow for growth-coupled evolution (Figure 4.3A).

At first, we deleted *edd*, encoding a 6-phosphogluconate dehydratase. This reaction is part of the ED pathway together with the sequential pyruvate-releasing step, encoded by *eda*, so the deletion of *edd* renders the whole pathway inactive. Next, we decoupled pyruvate from phosphoenolpyruvate (PEP), by deleting the pyruvate kinases encoded by *pyk* and *pykA*. The last major pyruvate node was removed by deleting phosphoenolpyruvate carboxylase, encoded by *ppc*. This reaction converts PEP into oxaloacetate, which subsequently can be converted to pyruvate by the oxaloacetate decarboxylase (PP_1389). It is important to note, that the deletion of *pyk*, *pykA*, and *ppc* not only disrupts the flow to pyruvate but also increases the intracellular PEP pool for the shikimate pathway. At last, we deleted *glpR*, which encodes the transcriptional repressor of the glycerol catabolic operon. We termed this strain Δ*pyr* and evaluated whether it exhibited pyruvate auxotrophy by growing it in glycerol minimal media supplemented with increasing pyruvate concentrations.

As expected, Δ*pyr* was not able to grow in glycerol minimal medium without the addition of pyruvate, indicating that biomass formation is exclusively dependent on its external supplementation (Figure 4.3B). Albeit this strain can serve as a base strain for ALE, our *in silico* simulations pointed to additional more favorable pyruvate-releasing reactions (AGPOP, CYSTL, SERD_L, ME2, OAADC, CYSDS, MCITL2, LDH_D, LDH_D2) that would allow faster growth rates than those achieved using CHRPL as pyruvate source (Figure 4.2A). Although the removal of these reactions would ben-

efit evolution, it might negatively affect cellular fitness. Moreover, the other chorismate-derived pyruvate-releasing reactions cannot be removed, as this would render the strain auxotrophic for both tryptophan and folate (Figure 4.2B). Therefore, by only selecting on growth, there is an increased possibility that undesired phenotypes develop.

To circumvent this problem, we implemented a second layer of screening by incorporating a previously established 4HB-responsive sensor [221]. This sensor constitutes a double mutant pobR enzyme from *Acinetobacter baylyi* ADP1, which binds to 4HB upon detection and expresses sfGFP. We further adapted this biosensor by including an LAA degradation tag, to decrease leakiness and increase tunability (Figure 4.3C). We cloned the sensor in a pSEVAb83 backbone and tested its efficacy in *P. putida* KT2440 in glycerol minimal media supplemented with increasing concentrations of 4HB. As expected, the sensor displayed dose-dependent *sfgfp* expression to exogenous 4HB concentrations indicating a strong selection method for ALE (Figure 4.3D). Using this method, superior strains exhibiting fast growth rates and a high flux towards 4HB formation can be isolated.

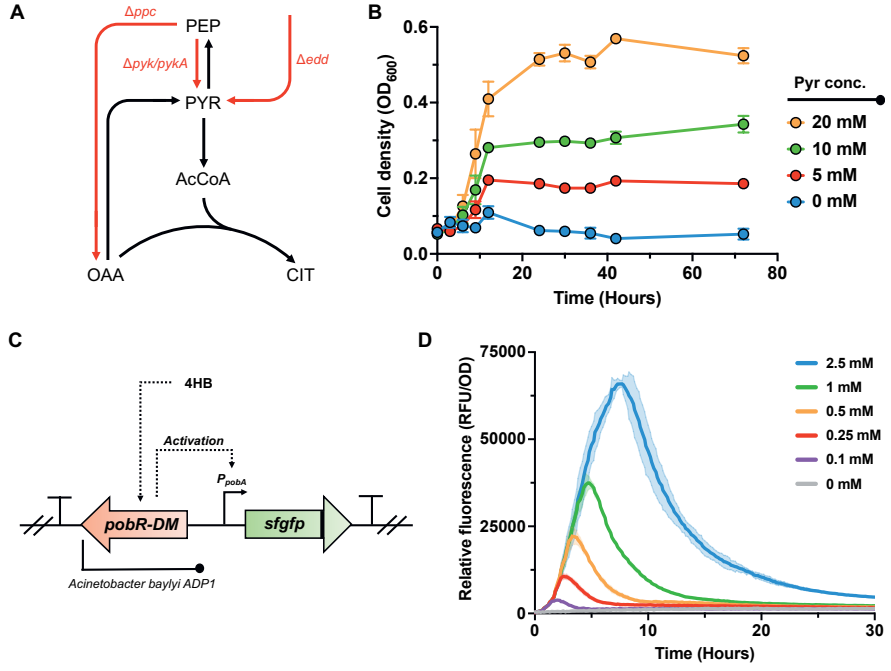


Figure 4.3: Base strain and biosensor for adaptive laboratory evolution. **A.** Metabolic scheme of the genetic basis of Δ*pyr*. All major pyruvate-releasing reactions that were deleted are denoted in red. **B.** Growth curve of Δ*pyr* in glycerol minimal media supplemented with various pyruvate concentrations. **C.** Graphical representation of the 4HB-responsive biosensor used in this study. The PobR enzyme from *Acinetobacter baylyi* ADP1 binds to 4HB upon its presence. This in turn activates P_{pobA} expressing the *sfgfp* gene. The gene is equipped with an LAA degradation tag, to reduce leakiness of the system. **D.** Relative fluorescence profiles of *P. putida* Δ*glpR* equipped with the 4HB-sensor upon exposure to increasing 4HB levels in glycerol minimal medium. Abbreviations: PEP, phosphoenolpyruvate; PYR, pyruvate; AcCoA, acetyl-CoA; OAA, oxaloacetate; CIT, citrate; 4HB, 4-hydroxybenzoate. Data points represent the mean value ± SD from three independent experiments.

Establishing SDC through laboratory evolution

Establishing SDC requires a rigorous rearrangement of the complete metabolic network. We established a pyruvate auxotrophic strain to make the shikimate pathway the major source for pyruvate and a 4HB-responsive sensor to select superior isolates. To further drive evolution, we incorporated a feedback-inhibition-resistant CHRPL [221]. To maintain high levels of CHRPL, we placed the corresponding gene (*ubiC*$^{E31Q/M34V}$) under the control of the constitutive J23100 promoter downstream of the biosensor in the same transcriptional operon. Through this design, the 4HB produced by CHRPL creates a positive feedback loop and increases its own transcription, aiding the evolution of SDC. We equipped *P. putida* Δ*pyr* with the modified biosensor and started evolution cultivating the strain in glycerol minimal medium in two independent experiments. Growth in the population emerged in the first phase after roughly 20-24 days in both experiments (Figure 4.4A). After this initial passage, growth rates quickly increased with the next passage, indicating that crucial adaptations occurred in the initial phase. In total, the cultures in the two independent experiments were serially diluted for ≈ 50 days in 13 to 18 passages and plated on minimal media with glycerol.

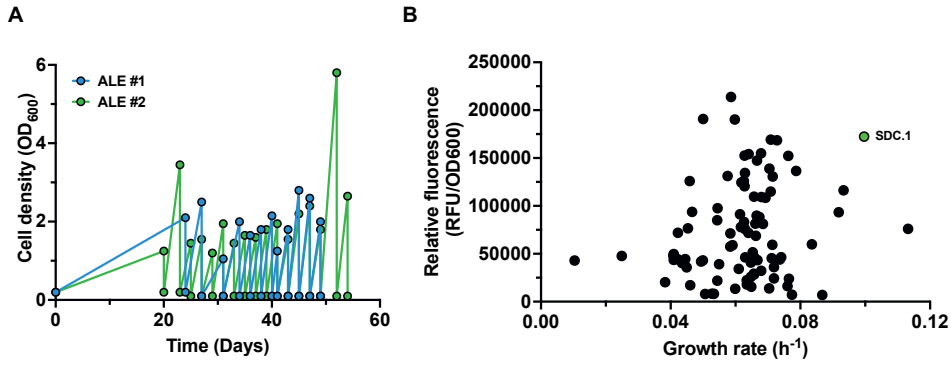


Figure 4.4: Isolation of SDC.1 after adaptive laboratory evolution (ALE) in glycerol minimal medium. **A.** Growth curves of ALE passages. **B.** Relative fluorescence profiles (RFU/OD600) plotted against the growth rate (h^{-1}) of the selected mutants after ALE. Every dot in the graph indicates a single isolated mutant.

Forty colonies were selected per ALE experiment for further characterization based on their fluorescence. The isolated mutants showed a diverse range of growth rates and fluorescence profiles (Sup. Figure 4.1A). Within the isolated population, we observed fast-growing mutants with low fluorescent profiles. Most likely, these strains evolved through alternative salvage pathways to restore growth. This finding highlights the importance of the implemented 4HB-responsive biosensor during evolution to select superior strains. From the heterogeneous mixture of mutants, we isolated SDC.1 (Figure 4.4B). This strain demonstrated fast growth among the isolated mutants (0.099 h^{-1}), albeit significantly slower than the wild type (0.210 h^{-1}) (Sup. Figure 4.1A). However, SDC.1 displayed high levels of fluorescence among the isolated mutants and a 10-fold increase in relative fluorescence compared to the wild type (Sup. Figure 4.1B). This indicates that the fast-growing mutant SDC.1 carries a high flux through the shikimate pathway towards 4HB biosynthesis.

Genomic characterization of SDC

To elucidate the genetic basis of SDC, we sequenced the genomes and plasmids of the most efficiently evolved strains. We focused solely on isolates that displayed high relative fluorescence. This is because these strains should display a high carbon flux through the shikimate pathway irrespective of the growth rate. From each evolution experiment, we chose ten mutants. At first, we sequenced the biosensor plasmids to check for alterations. However, no mutations occurred in the plasmids of all twenty isolates, indicating that growth occurred solely due to genomic alterations. All isolates from the two individual evolution experiments contained mutations in the *miaA* and *mexT* genes at various positions (Figure 4.5A).

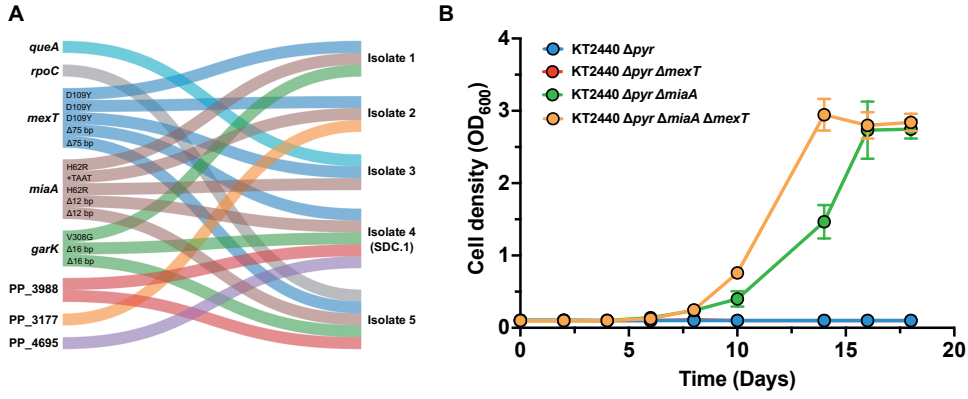


Figure 4.5: Exploring the genetic basis of SDC. **A.** Genomic alterations that were discovered in the five strains with the highest relative fluorescence after laboratory evolution. **B.** Growth curves of the reverse-engineered strains. The *miaA* and *mexT* genes were deleted separately and in combination from Δ*pyr*. Abbreviations: *queA*, S-adenosylmethionine:tRNA ribosyltransferase-isomerase; rpoC, DNA-directed RNA polymerase subunit beta; *mexT*, transcriptional regulator MexT; *miaA*, tRNA dimethylallyltransferase; *garK*, glycerate kinase. Data points represent the mean value ± SD from three independent experiments.

MiaA is a tRNA dimethylallyl transferase and has been known to affect the expression of various genes related to the central and secondary metabolism [222]. In *P. putida* specifically, the removal of *miaA* was reported to dramatically increased the expression of the *trpE* and *trpGDC* genes, both involved in tryptophan biosynthesis [223]. MiaA has been further studied in *Pseudomonas chlororaphis*, where its inactivation led to the up-regulation of the *trp* genes and *aroF*. The latter encodes a 3-deoxy-7-phosphoheptulonate synthase, which catalyzes the first reaction of the shikimate pathway [224]. The *mexT* gene encodes a transcriptional regulator that has mostly been studied in *Pseudomonas aeruginosa* in which it represses the entire quinolone biosynthetic pathway and the first reaction of tryptophan biosynthesis [225]. Another major mutation that occurred in 12 out of the 20 isolates was a perturbation in *garK*, encoding a glycerate kinase. However, as the *garK* perturbation did not occur in all isolates, we deemed this one as non-universal and focused solely on *miaA* and *mexT*.

To assess the importance of these mutations, we deleted the *miaA* and *mexT* genes from Δ*pyr* and assessed their growth in glycerol minimal medium. Surprisingly, the removal of the *miaA* gene allowed growth without the need for evolution (Figure 4.5B). The deletion of *mexT* did not result in immediate growth. However, in combination with Δ*miaA*, its deletion led to higher growth rates and a shorter lag phase than the single Δ*miaA* mutant. As the inactivation of only *miaA* allowed growth in minimal medium with glycerol in Δ*pyr*, we speculate that this gene is key in regulating the shikimate pathway and could be a potential target for metabolic engineering strategies in other organisms.

Evaluation of metabolic fluxes in SDC

Strain SDC.1 was selected from the heterogeneous mix based on its growth rate and fluorescence profile. Next, we aimed to determine the flux increase through the shikimate pathway of this strain compared to the wild type. Chorismate, the final product of the shikimate pathway, is equimolarly cleaved into 4HB and pyruvate by chorismate pyruvate lyase (CHRPL). Therefore, we set out to quantify 4HB production to evaluate whether SDC.1 uses the shikimate pathway as its main metabolic route for growth. In theory, the higher the yield of 4HB, the more flux is diverted into the shikimate pathway. FBA analysis predicts a maximum pathway yield of 0.39 mol/mol (4HB:glycerol). In this scenario, there is no bacterial growth, and all carbon, including the released pyruvate by CHRPL, is directed toward 4HB synthesis. Therefore, we set this as our attainable maximum. To assess 4HB production, we deleted the *pobA* gene in both the wild type and SDC.1, creating SDC.2. This gene encodes a p-hydroxybenzoate hydroxylase and is responsible for the degradation of 4HB to protocatechuate (PCA), which can be further degraded to fuel the TCA cycle. The removal of the *pobA* gene had a negligible effect on wild-type growth. However, the growth of SDC.2 was severely stunted, requiring 14 days to reach the stationary phase compared to 4 days for SDC.1 (Figure 4.6A). Moreover, a lower final cell density was observed indicating that SDC.2 metabolism is highly dependent on the activity of the shikimate pathway and that the produced 4HB cannot be recycled back for cell proliferation. The obtained 4HB yield for SDC.2 was 0.06 mol/mol, a 13.8–fold increase compared to the wild type Δ*pobA*, confirming a significantly increased carbon flux through the shikimate pathway (Figure 4.6B). However, this flux only comprises 15.4% of the predicted maximum theoretical yield, implying that other routes are still taking a significant portion of the intracellular fluxes.

We focused particularly on the shikimate pathway to further optimize SDC.2. We hypothesized that *quiC*, encoding 3-dehydroshikimate (3HDS) dehydratase, was siphoning off carbon from the shikimate pathway towards the TCA cycle. When growth was simulated for SDC.2, the model predicted 54% of the flux entering the shikimate pathway to be redirected through this reaction, converting the shikimate pathway intermediate 3-dehydroshikimate to protocatechuate, and efficiently circumventing the *pobA* deletion. As such, we created SDC.3 by deleting *quiC* from SDC.2 and assessed its growth and 4HB production. It became apparent that this reaction indeed was a major metabolic bypass. The SDC.3 strain displayed even further stunted growth compared

to SDC.2, reaching the stationary phase after 10 days with a concomitant significantly lower biomass formation (Figure 4.6A). The final 4HB yield increased 2.4–fold to 0.14 mol/mol (Figure 4.6D), indicating a further increase in the flux through the final reactions of the shikimate pathway.

Although the optimized strain SDC.3 demonstrated increased fluxes through the shikimate pathway, it only reached 36.7% of the attainable predicted maximum yield. Therefore, we further simulated the SDC metabolism to identify bottlenecks. Our simulations predicted that 19.7% of the total consumed glycerol was excreted as glycerate (Sup. Figure 4.2). As mentioned earlier, 12 out of 20 isolates contained mutations in the *garK* gene. The isolated SDC.1 and its derivatives contain a 16 bp deletion within the gene resulting in a frameshift. The *garK* gene encodes a glycerate kinase, which is responsible for the phosphorylation of glycerate to glycerate-2-phosphate, which can then enter the main metabolism. We hypothesized that the frameshift renders the enzyme inactive and allows glycerate to accumulate. We quantified glycerate production in SDC.3 and confirmed our hypothesis as 18.3% of the total amount of glycerol was excreted as glycerate, decreasing the total carbon flux towards the shikimate pathway (Figure 4.6C).

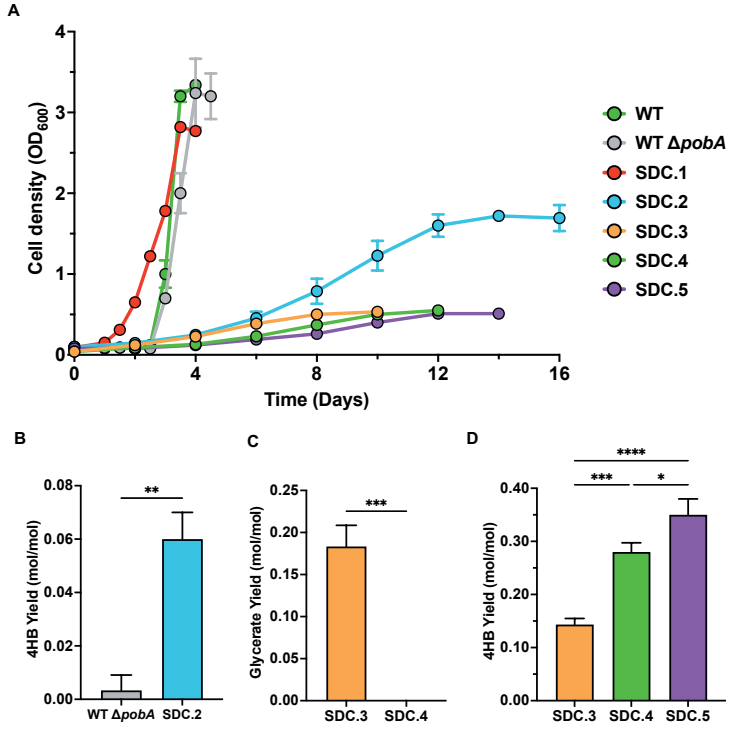


Figure 4.6: Characterization of SDC strains. **A.** Growth curves of the wild-type strain and WT Δ*pobA* compared to the SDC strains. **B.** Quantification of 4HB yields in the WT Δ*pobA* vs. SDC.2. strain. **C.** Quantification of glycerate yields in SDC.3 vs SDC.4. strains. **D.** Quantification of 4HB yields in SDC.3, SDC.4, and SDC.5 strains. Data points and bar graphs represent the mean value ± SD from three independent experiments. *, p <0.1; **, p <0.01; ***, p <0.001; ****, p <0.0001 determined by an unpaired Student's t-test. Note that all the strains (including WT) contain the pSensorEvo plasmid)

According to the model, pyruvate production from glycerol using SDC yields 6.1% of the ATP obtained in the ED pathway (Sup. Methods). Aerobic bacteria like *P. putida* generate most electron carriers in the TCA cycle, which then feed the electron transport chain and produce ATP through oxidative phosphorylation. However, in SDC, the connection between glycolysis and TCA is disrupted and the TCA cycle can only be reached through the ATP-consuming shikimate pathway. Model predictions show that in native metabolism, glycerol is catabolized to the glycolytic intermediate dihydroxyacetone-phosphate (DHAP) (Figure 4.1). This process costs 1 ATP per glycerol and yields 1 reduced quinone. However, in SDC, the model predicts NAD-dependent oxidation of glycerol to glycerate by AldB-1, PP_2694, or FrmA (Sup. Figure 4.2). Thus, the activation of the alternative glycerol utilization pathway likely serves as an additional source of reducing equivalents which can be oxidized in the electron transport chain to generate a proton motive force for additional ATP. In addition, the GarK enzyme also requires the utilization of ATP suggesting a relationship between *garK* mutation and energy conservation. Although the model predicts NAD-dependent oxidation of glycerol to glycerate, the first step in glycerate production from glycerol in *P. putida* is initiated by the PQQ-dependent alcohol dehydrogenases encoded by *pedE* and *pedH* [226]. Like NADH, $PQQH_2$ gets oxidized in the electron transport chain, generating ATP. The ability to generate ATP by converting glycerol into glycerate, therefore, reduces the required SDC-dependent flux towards the TCA cycle. To abolish glycerate production, we deleted *pedH* and *pedE* from SDC.3, creating SDC.4. This deletion extended the lag phase slightly but did not have a significant impact on the growth rate (Figure 4.6A). Glycerate production was completely abolished and revealed to be a major metabolic bottleneck as its removal increased the 4HB yield from 0.14 to 0.28 mol/mol (Figure 4.6D). This accounts for 71.0% of the maximum predicted pathway yield and indicates that a significant flux in SDC.4 is diverted through the shikimate pathway.

For further optimization, we focused on branching pathways from the shikimate pathway. Within the isolated mutants, we discovered mutations in the genes *miaA* and *mexT* and showed their impact on establishing SDC. Both their encoded proteins have been reported to directly influence tryptophan biosynthesis by regulating anthranilate synthase (ANS), encoded by *trpE* in *P. putida*. We hypothesized that the SDC.1 strain and its further derivatives may display increased ANS activity due to the mutations in *miaA* and *mexT*. Like CHRPL, the TrpE enzyme cleaves chorismate releasing pyruvate in the process. However, this reaction requires additional L-glutamine or ammonia as an amine donor, which according to our *in silico* assessment results in lower growth rates (Figure 4.2A). Yet, the removal of this reaction is unwanted as it would render the strain auxotrophic for tryptophan. Therefore, we aimed to reduce its activity by introducing a P290S point mutation in the *trpE* gene of SDC.4. This specific mutation has been reported to lower the activity of the TrpE enzyme [227]. This new strain, termed SDC.5, has a slower growth rate (0.008 ± 0.000 h^{-1}) compared to SDC.4 (0.011 ± 0.000 h^{-1}) yet reached similar final cell densities (Figure 4.6A). In this strain the 4HB yield on glycerol increased from 0.28 to 0.35 mol/mol (Figure 4.6D). This accounts for 89.2% of the maximum predicted pathway yield and indicates that SDC.5 diverts most of its glycerol metabolism through the shikimate pathway.

Genomic integration and characterization of a stable SDC strain

So far, we quantified the flux through the shikimate pathway using 4HB as output. By reaching the predicted maximum, we can conclude that all major bottlenecks are removed and that we successfully established SDC. Therefore, we aimed to produce a stable strain that could serve as a *chassis* for shikimate pathway-derived products. For this purpose, we integrated the feedback-resistant *ubiC*$^{E31Q/M34V}$ gene in SDC.5 under the control of a bicistronic design at the innocuous PP_5322 site [228]. Furthermore, we restored the *pobA* gene to create the final SDC.6 strain in which 4HB can once more fuel biomass production (Figure 4.7A). We determined the growth characteristics of SDC.6 and compared them to the wild type as both strains have different pathways to convert glycerol to pyruvate. Moreover, we used SDC.1 as a second control to examine the impact of all further modifications on growth. As expected, the wild type using the canonical glycerol metabolism grew fast and reached the stationary phase within 24 hours. Our genomically stable SDC.6 strain had a lower specific growth rate than SDC.1, 0.038 $\pm$ 0.001 h^{-1} vs 0.052 $\pm$ 0.007 h^{-1}, respectively (Figure 4.7B). We hypothesize that the lower growth rate of SDC.6 is the result of i) the absence of QuiC, which allowed a quick bypass in SDC.1 from the shikimate pathway toward the TCA cycle, ii) the removal of glycerate production, which allowed higher ATP generation, and iii) fewer pyruvate production due to decreased TrpE activity (Figure 4.7A).

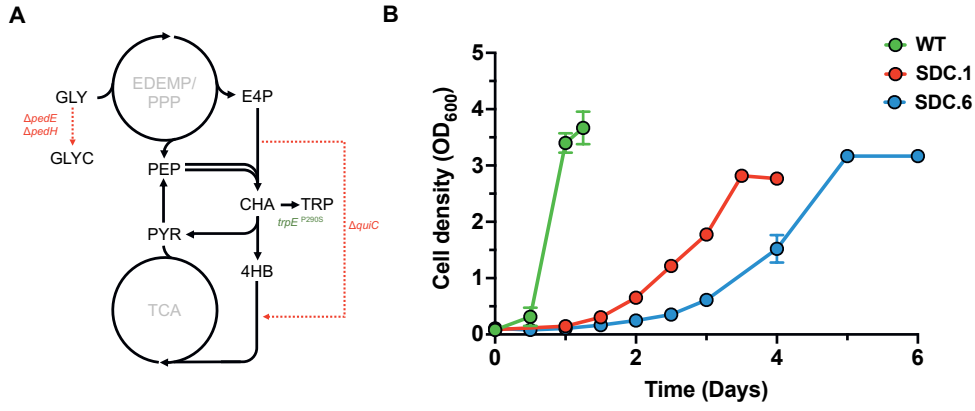


Figure 4.7: Characterization of the stable SDC strain SDC.6. **A.** Graphical representation of the genetic modifications in SDC6. The *pedE* and *pedH* genes were deleted to abolish glycerate formation as a byproduct. The *quiC* gene was deleted to avoid splitting of the shikimate pathway and the P290S point mutation in *trpE* was introduced to lower fluxes toward tryptophan biosynthesis. **B.** Growth curves of the wild type (KT2440), the first isolated SDC.1 and optimized SDC.6 strain. Abbreviations: GLY, glycerol; GLYC, glycerate; E4P, erythrose-4-phosphate; PEP phosphoenolpyruvate; PYR, pyruvate; CHA, chorismate; TRP, tryptophan; 4HB, 4-hydroxybenzoate; EDEMP, Entner-Doudoroff-Embden-Meyerhof-Parnas pathway; PPP, pentose phosphate pathway; TCA, tricarboxylic acid cycle. Data points represent the mean value $\pm$ SD from three independent experiments. Note that only strain SDC.1 contains the pSensorEvo plasmid.

Although slower, SDC.6 reached a final OD600 of 3.17, which is 14.5% higher than SDC.1. This can be attributed to the removal of glycerate formation in SDC.6 which can now be funneled towards biomass production (Figure 4.7A). However, this removal likely increased the lag phase in SDC.6. The PedE and PedH enzymes are part of the alternative glycerol metabolism in *P. putida* and their removal has been demonstrated to strongly prolong the lag phase [226]. Yet, SDC.6 is still superior to SDC.1 as it is devoid of all the bottlenecks described above and can therefore lead to higher yields for shikimate pathway-derived products.

Next, we used the model to calculate the maximum specific growth rate in SDC.6. For this analysis, we set the maximum glycerol uptake rate at 3.95 mmol/g_{DW}/h, the same as in KT2440. However, the model predicts a lower glycerol uptake rate for this strain, with a maximum of 3.36 mmol/g_{DW}/h, probably indicating a limitation in cofactor regeneration. Moreover, the model predicts a maximum specific growth rate of 0.16 h^{-1}, which is 4.2–fold higher than the actual growth rate we observed. Therefore, there is still room for improvement to establish SDC.6 as a true *chassis* for shikimate pathway-derived products.

Discussion

In this study, we created a new-to-nature shikimate pathway-dependent catabolism in *P. putida*. Based on model simulations and rational engineering, we constructed a growth-coupled scenario. Next, through pyruvate-driven evolution coupled with a biosensor-assisted selection strategy, the superior mutant SDC.1 was isolated. This strain showed a fast growth rate and a high relative fluorescence profile among all isolates. We discovered that SDC.1 displayed a 13.8-fold higher flux through the shikimate pathway than the wild type, reaching a 4HB yield of 15.2% of the maximum predicted pathway yield. Through further model-driven and rational engineering, we removed all potential bottlenecks and created SDC.5 which reached 89.2% of the maximum predicted pathway yield, demonstrating a massive catabolic and anabolic flux through the shikimate pathway. By reintroducing 4HB degradation, we established SDC.6, the first strain ever constructed that uses the shikimate pathway as the dominant pathway for glycerol catabolism during growth. Moreover, the SDC.5 strain presented in this study can be equipped with other pyruvate-releasing steps to achieve high yields for valuable molecules such as maleate [229], 3-hydroxybenzoate [230], p-aminobenzoate [231], and salicylate [232].

Growth-coupled bioproduction of shikimate-derived products using pyruvate-releasing reactions has been theorized before [202] and has recently also been proven [232, 233]. However, although both strategies were effective, they still relied on the external supplementation of aromatic amino acids and yeast extract, which would substantially increase operating costs. The fact that the carbon flux in these designs could not support growth could potentially be attributed to the complex regulatory network of the shikimate pathway. This pathway is regulated on different levels *e.g.*, transcriptional repression, attenuation, and feedback inhibition [234], which has made traditional metabolic engineering strategies rather challenging. In this study, we showcase that

indeed regulation is the biggest bottleneck. The SDC was established due to a small number of mutations that most likely broke the regulatory network of the shikimate pathway. Various genomic alterations were discovered in all sequenced isolates in the genes *miaA* and *mexT*. Both encoded enzymes appear to have a regulatory function regarding the shikimate pathway, yet their exact regulatory mechanism remains elusive. Deletion of *miaA* was able to restore growth without the need for evolution and the subsequent deletion of *mexT* further improved growth rates and reduced lag phases. Therefore, we believe that it is noteworthy to examine these genes as metabolic engineering targets for the shikimate pathway in other industrial organisms.

In this work, FBA analysis was frequently used to guide and support the generation of hypotheses. One noticeable example was the prediction of glycerate as a by-product. In SDC.1 this phenotype was attributed to a 16 bp deletion in the *garK* gene, encoding a glycerate kinase. This enzyme is responsible for the phosphorylation of glycerate to glycerate-2-phosphate, which can then enter the main metabolism. The production of glycerate from glycerol produces two molecules of $PQQH_2$ [226], which yields two molecules of ATP in the electron transport chain. However, the GarK enzyme requires the usage of one molecule of ATP. As SDC is energetically poor, this deletion likely inactivated the enzyme to conserve ATP, which allowed glycerate to accumulate. Moreover, glycerate production results in the generation of reduced cofactors which can be used for respiration, decreasing the required catabolic flux over the SDC and supporting the hypothesis of ATP generation as the main limiting factor in SDC. Although the pyruvate released by CHRPL can be oxidized in the TCA cycle to generate reducing equivalents for respiration, the high yields observed in the SDC.5 strain suggest the recycling of pyruvate to PEP to further fuel 4HB production. When using SDC, this behavior would promote product formation but results in low growth rates that could limit productivity.

To promote growth in SDC, external electron donors could be applied to generate NADH, which then can be oxidized in the electron transport chain to provide ATP without carbon loss. Formate is a formidable candidate as *P. putida* already encodes a kinetically fast formate dehydrogenase [235]. Another potential electron donor is phosphite, whose potential has recently been explored in *P. putida* [236]. Although the phosphite dehydrogenase is kinetically slower than its formate counterpart [237], it was able to increase the NADH pool in *P. putida*. Moreover, compared to formate, phosphite metabolism gives a competitive advantage, allowing non-sterile fermentation conditions, and lowering overall production costs [238]. Therefore, both electron donors are worthy of investigation to further improve the SDC.

In our current design, the shikimate pathway carries the whole metabolic regime with a single overexpression. Thus, further optimization of this pathway is a necessity to increase the fluxes. This could either be achieved through genomic overexpression or a second round of laboratory evolution. In our first evolution experiment, we identified *miaA* and *mexT* as key players involved in the shikimate pathway. However, the strain still had many bottlenecks that, although facilitated growth, decreased the shikimate pathway flux. The streamlined SDC.5 strain has a maximized flux through the shikimate pathway where the released pyruvate is the only junction between glycerol

catabolism and the TCA cycle. This strain would be an excellent candidate for another round of evolution. Not only would this stringent selection system lead to increased fluxes through the shikimate pathway, but it could also lead to the identification of other unknown bottlenecks.

Additionally, fluxes in SDC could be optimized through rational, model-guided engineering. We demonstrated the power of *in silico* simulations to aid the experimental design, identify metabolic bottlenecks, and guide optimization based on maximum yields and optimal growth rates. Compared to the wild type, SDC strains have a complicated glycerol-to-pyruvate node with more enzymatic steps that make rational engineering nontrivial. The integration of metabolic modeling and metabolomic analysis could facilitate the evaluation of new model-guided optimization approaches [239]. This optimization could focus on the balance between the shikimate pathway precursors, erythrose-4-phosphate (E4P) and phosphoenolpyruvate (PEP). Besides low energy production, we hypothesize the availability of E4P as an additional limitation of SDC, as the PEP pool is already significantly increased through the blockage of its degradation nodes in the initial Δ*pyr* strain. Like the shikimate pathway, the pentose phosphate pathway is predominantly used for anabolic reactions in *P. putida*, likely displaying low native fluxes [240]. A model-driven strategy could be applied to improve growth of SDC.6, selecting overexpression targets that, while maximizing growth, minimize resource allocation towards the expression of unnecessary enzymes [241].

This work highlights the plasticity of bacterial metabolism and how a combination of model-driven design, rational engineering, and laboratory evolution can create novel metabolisms. We repurposed the shikimate route as the major catabolic route and demonstrated that it can carry the whole cellular flux for growth. Moreover, we believe that the SDC strain presented in this study will open many potential practical applications for the high-yield synthesis of industrial valuable aromatic compounds.

Declaration of interest

Vitor A. P. Martins dos Santos has interests in LifeGlimmer GmbH.

Acknowledgment

This project was founded by the European Union Horizon2020 projects EmPowerPutida (grant number 635536) and IBISBA (grant numbers 730976 and 871118), and the NWO (project number GSGT.2019.008).

Data availability

Supplementary tables are available at Zenodo and supplementary methods and figures are presented below.

Supplementary Methods

Energy generation by SDC

We calculated overall equations from glycerol to pyruvate to study the feasibility of SDC in terms of its ability to generate reduced cofactors and ATP compared to native metabolism. Native and SDC metabolism were simulated with flux balance analysis (FBA) with pyruvate production (EX_pyr_e) as maximization objective. Then, reduced models were generated containing only predicted active reactions of central carbon metabolism (Sup. Table 5), as well as exchange reactions for NADH (nadh_c), NADPH (nadph_c), ATP (atp_c), NAD (nad_c), NADP (nadp_c), pyruvate (pyr_c), H^+ (h_c), H_2O (h2o_c), ubiquinone-8 (q8_c), ubiquinol (q8h2_c), oxygen (o2_c), coenzyme A (coa_c), CO_2 (co2_c) and inorganic phosphate (pi_c) in the cytoplasm. To find the most efficient overall reaction in terms of ATP generation, pyruvate exchange was set as the objective to maximize, and the bounds of its exchange reaction were constrained to the optimal value. Then, the ATP exchange reaction was set as a new objective to maximize. A glycerol uptake of 1 mmol/g_{DW}/h was used and the fluxes through the different exchange reactions are defined as coefficients in the overall equations. According to iJN1462 oxidation of NADH via NADH dehydrogenase (reaction NADH16pp) allows the export of 3 H^+ to the periplasm and the generation of ubiquinol (q8h2). The reduction of ubiquinone-8 (q8) to q8h2, and oxidation of q8h2 by cytochrome oxidase (CYTBo3_4pp) allows the export of 4 H^+ to the periplasm. Generation of ATP by the ATP synthase (ATPS4rpp) requires the import of 4 H^+ from the cytoplasm resulting in a yield of 1 ATP per oxidation of 1 q8h2 and 1.75 ATP per oxidation of NADH.

Overall equation native metabolism (KT2440):

$1Glycerol + 1Q8 + 1NAD^+ + 1ATP + P_i \rightarrow 1PYR + 1Q8H_2 + 1NADH + H^+ + 2ATP$

Simplified equation (conversion of reduced cofactors to ATP in the respiratory chain):

$1Glycerol \rightarrow 1PYR + 3.75ATP$

Overall equation SDC:

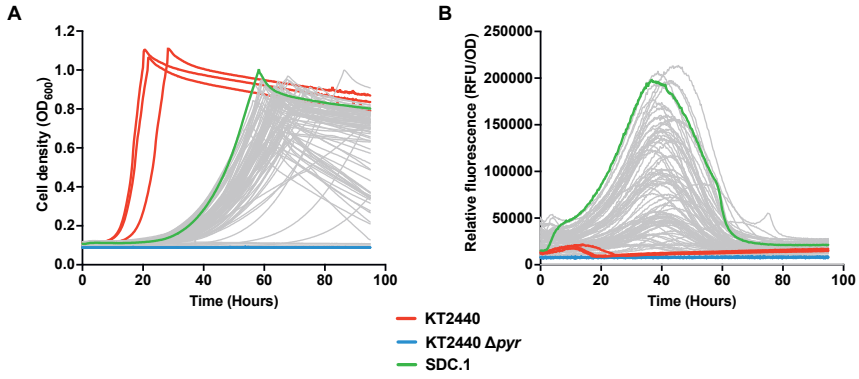
$1Glycerol + 1.8Q8 + 1.4NAD^+ + 1.1NADPH + 1.1O_2 + 2.1ATP + 1.3H_2O \rightarrow 0.55PYR + 1.8Q8H_2 + 1.4NADH + 1.4CO_2 + 2.1ADP + 2.1P_i + 2.9H^+$

Simplified equation (assuming conversion of NADH to NADPH and generation of ATP by oxidation of reduced cofactors in the respiratory chain):

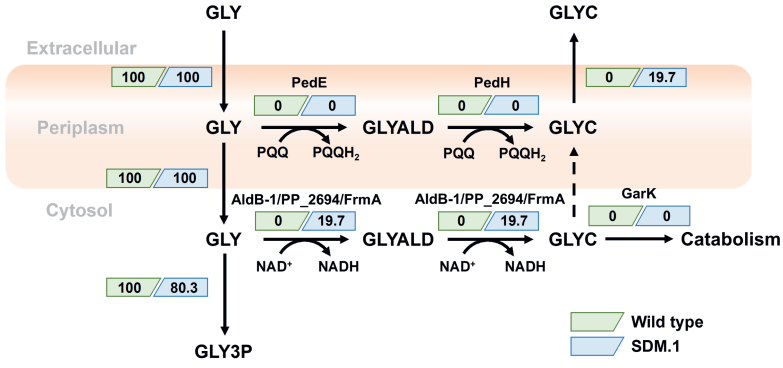
$1Glycerol + 1.1O_2 \rightarrow 0.55PYR + 1.4CO_2 + 0.23ATP$

Note: the O_2 in this equation is not used in the respiratory chain but in the first two reactions converting 4HB into TCA intermediates.

Supplementary Figures



Sup. Figure 4.1: Characterization of isolated mutants after adaptive laboratory evolution. **A.** Growth curves of selected mutants after ALE in glycerol minimal medium. **B.** Relative fluorescence profiles of the selected mutants after ALE in glycerol minimal medium. All strains in the depicted graphs are equipped with the 4HB responsive biosensor, and the evolved mutants were compared to the wild type (red) and the parental Δ*pyr* strain (blue). Every line indicates a single strain.



Sup. Figure 4.2: Model predictions of wild type and SDM.1 strain fluxes (%) in the conversion of glycerol to glycerate. Abbreviations: GLY, glycerol; GLYALD, glyceraldehyde; GLYC, glycerate; AldB-1, aldehyde dehydrogenase; FrmA, glutathione-dependent formaldehyde dehydrogenase; PedE/H, PQQ - dependent alcohol dehydrogenases; GarK, glycerate kinase.

CHAPTER 5

Enzyme-constrained models predict the dynamics of *Saccharomyces cerevisiae* growth in continuous, batch, and fed-batch bioreactors

Sara Moreno Paz, Joep Schmitz, Vitor A. P. Martins dos Santos, María Suárez Diez

This chapter is published in *Microbial Biotechnology*
doi.org/10.1111/1751-7915.13995

Abstract

Genome-scale, constraint-based models (GEM) and their derivatives are commonly used to model and gain insights into microbial metabolism. Often, however, their accuracy and predictive power are limited and enable only approximate designs. To improve their usefulness for strain and bio-process design, we studied their capacity to accurately predict metabolic changes in response to operational conditions in a bioreactor, as well as intracellular, active reactions. We used flux balance analysis (FBA) and dynamic FBA (dFBA) to predict the growth dynamics of the model organism *Saccharomyces cerevisiae* under different industrially relevant conditions. We compared simulations with the latest developed GEM for this organism (Yeast8) and its enzyme-constrained version (ecYeast8) herein described with experimental data and found that ecYeast8 outperforms Yeast8 in all the simulations. EcYeast8 was able to predict well-known traits of yeast metabolism including the onset of the Crabtree effect, the order of substrate consumption during mixed carbon cultivation, and the production of a target metabolite. We showed how the combination of ecGEM and dFBA links reactor operation and genetic modifications to flux predictions, enabling the prediction of yields and productivities of different strains and (dynamic) production processes. Additionally, we present flux sampling as a tool to analyze flux predictions of ecGEM, of major importance for strain design applications. We showed that constraining protein availability substantially improves the accuracy of the description of the metabolic state of the cell under dynamic conditions. This therefore enables more realistic and faithful designs of industrially relevant cell-based processes and, thus, the usefulness of such models.

Introduction

One of the goals of biotechnology is the design of cell factories to produce metabolites of industrial interest. Metabolic engineering introduces heterologous pathways and rewires cell metabolism to increase product yield, titer, and productivity [242]. However, although the production capacity of microorganisms is affected by many external factors such as oxygen and carbon availability, these interactions are often underestimated during the strain design process. The lack of a strong link between initial strain design and industrial deployment causes the so-called "Valley of Death", where only one in 5,000 to 10,000 innovations make the long route from initial finding to market implementation [88, 243, 244]. Models of microbial metabolism are increasingly used to aid the design and steering of bio-processes in an attempt to navigate the "Valley of Death". We studied the capacity of these models to provide accurate predictions of intracellular active fluxes, key to guiding metabolic engineering strategies. Besides, we tested their ability to link strain and bio-process design (*i.e.* how modifications in the reactor environment impact predictions on cell metabolism).

Genome-scale metabolic models (GEM) are mathematical representations of cell metabolism able to establish genotype-phenotype relationships linking genes and enzymes with metabolic reactions. These models are based on annotated genomes and can be expanded to include resource allocation constraints such as maximum membrane surface area or cell volume [66]. Sánchez et al. introduced the GEM with Enzymatic Constraints using Kinetic and Omics (GECKO) framework to generate enzyme-constrained models (ecGEM) by adding additional constraints linked to the limited enzyme production capacity of the cell [67]. In these models, protein abundance and enzyme turnover values (k_{cat}) limit the flux of the corresponding reactions. The ecGEM of *Saccharomyces cerevisiae* enables a more extensive and accurate simulation of microbial physiology including overflow metabolism, stress responses, and consumption rates of different carbon sources.

Flux balance analysis (FBA) is the most common method to simulate genome-scale metabolism. It uses linear programming to optimize an objective function and has extensively been used to predict cellular growth, and product secretion patterns and to develop overproduction strains [245, 246, 247, 248]. FBA assumes time-invariant extracellular conditions consistent with chemostat operation. Still, industrial-scale production is often achieved with batch and/or fed-batch cultures where extracellular conditions vary in time. Therefore, dynamic FBA (dFBA) extends FBA by introducing kinetic equations for extracellular metabolites and biomass. dFBA has been applied to simulate *Escherichia coli* industrial fermentations, compare ethanol production of different *Saccharomyces cerevisiae* strains during fed-batch growth, and identify industrially relevant bottlenecks for ethanol production from xylose [249, 250, 251]. Whereas FBA only captures one of the multiple solutions that leads to the optimization of the desired objective, sampling algorithms provide distributions of feasible flux solutions that represent the whole feasible flux space. Besides, the establishment of an objective function, which may introduce bias in the predictions, is not required [64].

We used FBA and dFBA to predict the growth dynamics of *S. cerevisiae* under industrially relevant conditions and compared simulations using Yeast8 (GEM) and ecYeast8 (ecGEM) with experimental data. We challenged the models to predict changes in cell metabolism (substrate uptake, growth, and product secretions) in response to the operation of the reactor, constituting one of the few examples of the combination of ecGEM and dFBA. For the first time, we used flux sampling of ecYeast8 to evaluate central carbon metabolic fluxes at a range of growth rates representative of chemostat, fed-batch, and batch growth of *S. cerevisiae*. We tested how flux sampling can be used to study central metabolic fluxes, of major importance for strain design applications. We provide a set of scripts to easily implement dFBA on traditional and ecGEM as well as a validation dataset containing fermentation-related data of *S. cerevisiae* cells growing in chemostat, batch, and fed-batch reactors. We show how the combination of ecGEM, dFBA, and flux sampling enables more realistic and faithful designs of industrially relevant cell-based processes and, thus, increases the usefulness of such models.

Materials and methods

Yeast8 and ecYeast8 models were obtained from Lu et al. [166] and, unless stated differently, default values for upper and lower bounds of reactions were used during the simulations. K_{cat} values in ecYeast8 were rescaled and additional constraints were imposed (see Sup. Methods). Model simulations were performed using Python 3.6, COBRApy (v0.18.1), and glpk as solver [163]. For details on experimental data used in this study see Sup. Methods and Sup. Table 5.1. Functions developed for chemostat, batch, and fed-batch simulations as well as an example of their use are available at Gitlab.

Glucose-limited chemostat simulations

During chemostat simulations, metabolic fluxes were calculated setting the bounds of the biomass reaction (r_2111) equal to the dilution rate (D, h^{-1}) and minimizing glucose consumption as objective for FBA optimization (maximizing r_1714 for Yeast8 and minimizing r_1714_REV for ecYeast8) [63]. The dilution rate was varied from 0.05 h^{-1} to 0.42 h^{-1} in intervals of 0.02 h^{-1} and feeds with glucose concentrations of 5 g/l, 7.5 g/l, 10 g/l, 15 g/l, and 30 g/l were simulated. In all cases the simulated cultures were glucose-limited, there was negligible glucose accumulation in the media, and the glucose mass balance was used to calculate the cell concentration in the reactor. If by-product secretion was predicted during simulations, their concentration was calculated using mass balances. A detailed explanation of the equations used is shown in the Sup. Methods.

Batch and fed-batch simulations

In batch and fed-batch simulations the growth reaction (r_2111) was set as the objective to maximize [63]. Following Sánchez et al., the upper bound of the protein pool reaction was increased by 25% [67]. When ethanol was present in the reactor, its uptake was allowed removing constraints on reactions r_1761 (Yeast8) or r_1761_REV (ecYeast8).

During the simulation of batch and the batch phase of fed-batch reactors, the glucose exchange reaction (r_1714 or r_1714_REV) was constrained based on the glucose concentration in the reactor using a Michaelis-Menten kinetic equation ($q_{glc,\,max}$ = 10 mmol/g_{DW} /h, $k_{m,glc}$ = 0.28 mM [252]). The glucose mass balance was used to calculate the remaining glucose in the reactor. During the feeding phase of fed-batch reactors, the glucose mass balance was used to calculate the glucose uptake rate and constrain the glucose exchange reaction. During this phase glucose is the limiting factor and its concentration in the reactor is negligible. After FBA optimization, the predicted metabolic fluxes were used to calculate new cell and by-product concentrations in the reactor using integrated mass balances. See the Sup. Methods for a detailed explanation of the equations used.

In simulations of batch growth on combinations of sucrose and glucose, sucrose and fructose, and sucrose and mannose, the uptake of glucose, fructose, mannose and sucrose was allowed if these metabolites were present in the reactor by setting a negative lower bound (Yeast8) or a positive upper bound (ecYeast8) to their exchange reactions (r_1714 and r_1714_REV, r_1709 and r_1709_REV, r_1715 and r_1715REV, r_2058 and r_2058_REV, respectively). This upper bound was calculated using a Michaelis-Menten equation for glucose. For the rest of the substrates a lower bound of -10 mmol/g_{DW}/h was used in simulations with Yeast8 and the upper bound of these reactions was unconstrained in ecYeast8. Additional simulations with ecYeast8 were performed including specific constraints based on inhibition of substrate uptake by some of the carbon sources (see Sup.Methods).

Simulation of a Δ*pdc* lactate producing *S. cerevisiae* strain

Yeast8 and ecYeast8 were modified to simulate a strain without a pyruvate decarboxylase (PCD) activity and expressing the lactate dehydrogenase gene from *Lactobacillus plantarum* [253]. In both models the growth reaction (r_2111) upper bound was constrained to 0.13 h^{-1} to simulate the maximum growth rate observed experimentally [253]. dFBA was used to simulate cells growing in a 1 l reactor operated as a batch with 100 g/l of initial glucose and a 100 g glucose pulse 75 hours after inoculation. Oxygen limitation was experimentally observed from 24 h after inoculation until the end of the process and was simulated constraining the oxygen exchange reaction (r_1992 in Yeast8 and r_1992_REV in ecYeast8) assuming no oxygen accumulation [253]. During simulations with ecYeast8, the export of products different than biomass, lactate, succinate, and glycerol was avoided constraining their secretion reactions [253]. The export of metabolites was unconstrained in simulations with Yeast8 to avoid infeasible solutions. See the Sup. Methods for details on the simulations.

Sampling of intracellular fluxes

The Artificial Centering Hit-and-Run (ACHR) Sampler from cobrapy was used to sample the solution space. Before sampling, the bounds of the biomass reaction were constrained to the desired growth rate, glucose uptake was set as an objective to minimize and the flux through this reaction was constrained to the minimal flux $\pm$ 10%. In all cases, the modified model was used and 10,000 samples were taken (see Sup. Methods). Samples that contained fluxes that violated lower and/or upper bounds or the steady state assumption were discarded using achr.validate [163]. Samples were taken at a range of growth rates (0.01, 0.05, 0.1, 0.15, 0.20, 0.23, 0.24, 0.25, 0.26, 0.27, 0.28, 0.29 and 0.3 h^{-1}). To analyze intracellular fluxes the median $\pm$ the median absolute deviation (MAD) of the valid samples was considered as the predicted flux through a given reaction. Sampling data is available in Gitlab.

Results

Chemostat simulations

S. cerevisiae cells grown in continuous cultures change their metabolism depending on the dilution rate. At low growth rates, they present a completely aerobic metabolism whereas ethanol production is observed at growth rates higher than the critical dilution rate (D_{crit}), a process known as the Crabtree effect. Data from chemostat growth of *S. cerevisiae* strains CBS8066, DS28911, and H1022 was obtained from literature [254, 255, 256]. In these experiments, *S. cerevisiae* was grown at different dilution rates (D) with different glucose concentrations in the feed. For each dilution rate, we constrained the growth rate of Yeast8 and ecYeast8 and used mass balances to calculate the cell, glucose, and by-product concentrations in the reactors at steady state.

Figure 5.1A shows predictions of biomass concentrations by Yeast8 and ecYeast8. Whereas Yeast8 predicts constant biomass concentration, ecYeast8 simulates a decrease in biomass concentration after a specific dilution rate, the critical dilution rate. The decrease in biomass concentration is also observed in the experimental data, which shows different critical dilution rates for different strains [257]. The model predicts a critical dilution rate of 0.27 h^{-1}, in agreement with that reported for strains DS28911 and H1022 (0.28 h^{-1} and 0.21 h^{-1}) [255, 256]. Strain CBS8066 has a higher protein content than H1022 and shows a higher critical dilution rate (0.38 h^{-1}) [254, 258]. This higher growth rate was simulated increasing protein availability in the model (26.8% increase of the upper bound of the protein pool reaction) showing that tuning protein availability results in different D_{crit}, suitable to predict chemostat growth of different *S. cerevisiae* strains.

Figure 5.1 also shows the maximum growth rate predicted by the model with the default bound for the protein pool reaction is 0.30 h^{-1} (0.38 h^{-1} when this bound is increased) while all strains can grow at dilution rates as high as 0.4 h^{-1}. However, when cells are grown experimentally at dilution rates higher than 0.3 h^{-1}, the dilution rate has to increase in small steps to avoid washout, indicating cells need time to adapt to high growth rates [256]. This adaptation is related to an increase in

protein content and therefore, chemostat predictions at high growth rates would improve with a growth rate-dependent protein availability constraint. Interestingly, decreasing maintenance requirements in the model did not affect maximum growth rate predictions, suggesting that the protein availability constraint implicitly accounts for protein synthesis costs and reduces the impact of the maintenance reaction in the simulations.

According to simulations with Yeast8, specific glucose uptake is proportional to the dilution rate. However, experimental data and simulations with ecYeast8 show a sharp increase in glucose uptake after D_{crit} (Figure 5.1B). Higher glucose uptake rates and lower biomass concentrations result in a decrease in the biomass yield on glucose after D_{crit}, which is only predicted by simulations using ecYeast8 (Figure 5.1C). Similar to the glucose uptake rate predictions by Yeast8, oxygen uptake rates, and CO_2 production rates are predicted to be proportional to the growth rate (Figures 5.1D,E). However, after D_{crit} cells show a partially fermentative metabolism that results in a decrease in the oxygen uptake rate and an increase in the CO_2 production rate. Besides, ecYeast8 predicts by-product formation at growth rates higher than the critical dilution rate. It predicts the secretion of acetaldehyde and acetate and accurately predicts ethanol flux at different glucose uptake rates (Figure 5.1F). None of these changes are predicted by Yeast8.

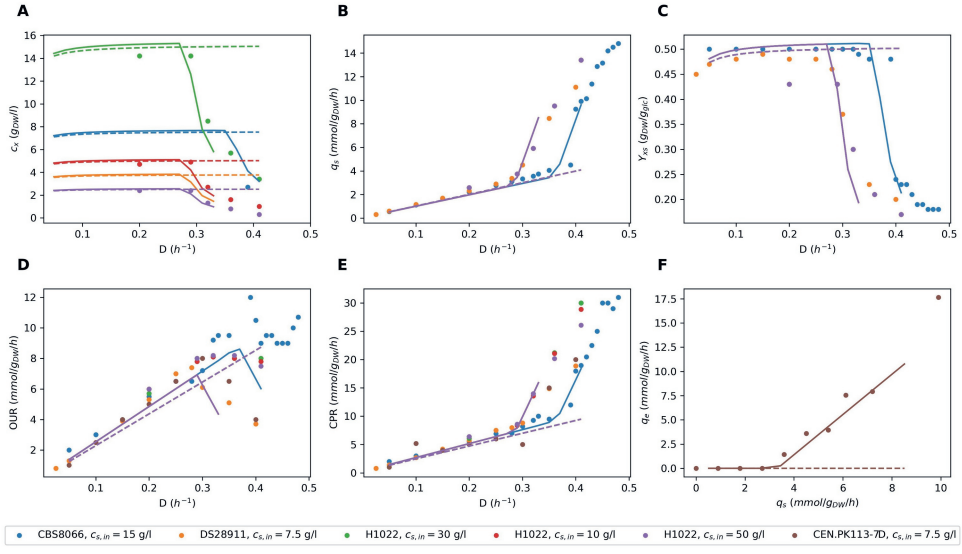


Figure 5.1: Chemostat simulations with Yeast8 (- -) and ecYeast8 (-) compared with experimental data (symbols) at different dilution rates (D) (**A-E**) or different specific glucose uptake rates (q_s) (**F**). **A.** Biomass concentration (c_x). **B.** Specific glucose uptake rate (q_s). **C.** Yield of biomass on glucose (Y_{xs}). **D.** Specific oxygen uptake rate (OUR). **E.** Specific CO_2 production rate (CPR). **F.** Specific ethanol production rate (q_e). Experimental data of strains CBS8066, DS28911, H1022, and CEN.PK 113.7D where obtained from [254, 255, 256, 259] respectively. Note that in figures B-F all dashed lines overlap and continuous orange, green, and red lines overlap with the continuous purple line.

Batch and fed-batch simulations

During batch fermentations glucose is present in excess and cells grow at their maximum growth rate. Yeast8 and ecYeast8 were used to simulate batch growth of *S. cerevisiae* and the results were compared to experimental data [260]. The glucose uptake rate was constrained in both models as a function of the glucose concentration in the reactor according to Michaelis-Menten kinetics. Whilst glucose uptake was the only constraint imposed on Yeast8, ecYeast8 was also limited by the availability of proteins.

Simulations using Yeast8 predict no ethanol production, faster glucose consumption, and higher cell concentrations than the experimental measurements (Figure 5.2). In these simulations, glucose uptake kinetics determines how fast glucose is consumed and all fluxes are distributed to optimize biomass production which results in exponential growth, no by-product formation as well as glucose depletion and growth arrest after 5h. Contrarily, simulations using ecYeast8 accurately predict glucose and biomass concentrations until glucose is depleted 7 h after inoculation. This model also predicts the production of ethanol and its consumption after glucose depletion. During these simulations, the growth rate is limited by protein availability, and only at glucose concentrations approaching k_m (0.28 mmol/l), the Michaelis-Menten equation for glucose uptake becomes the limiting factor. The protein availability constraint results in ethanol production by ecYeast8 and a realistic yield of biomass on glucose, overestimated by Yeast8. Although ethanol consumption was allowed during the entire simulation, it was only predicted after glucose depletion (in agreement with experimental data). However, during this phase, ecYeast8 simulates higher biomass concentration than the experimental measurements.

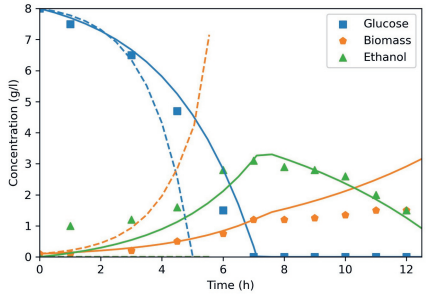


Figure 5.2: Batch simulation of *S. cerevisiae* H1022 with Yeast8 (- -) and ecYeast8 (-) compared with experimental data (symbols) [260].

In fed-batch reactors, batch growth is followed by a feeding phase in which media with substrate enters the reactor. During this phase, cellular growth is determined by the available glucose. We performed fed-batch cultivation of *S. cerevisiae* CEN PK-113-7D and used oxygen uptake and CO_2 production rates (OUR, CPR) as an indication of cell metabolism. This process was simulated using dFBA and model predictions were compared to experimental data. Yeast8 showed higher OUR and CPR as a result of a higher growth rate during the batch phase which

resulted in infeasible solutions before the start of the feeding phase. Simulations with ecYeast8 resulted in accurate prediction of OUR, CPR, and biomass concentration in the reactor (Sup. Figure 5.1).

Batch growth on multiple carbon sources

Yeast8 and ecYeast8 were used to simulate batch growth of *S. cerevisiae* in a mixture of carbon sources using dFBA. Dynesen et al. combined sucrose, a disaccharide of glucose and fructose, with glucose, fructose, or mannose to study growth and catabolite repression of *S. cerevisiae* DGI342 [261]. Simulations using Yeast8 and ecYeast8 were compared with this experimental data.

Yeast8 predicts simultaneous consumption of all carbon sources and unrealistically high uptake rates resulting in substrate depletion after 6 hours (Figure 5.3 A-C). In order to obtain better predictions, uptake reactions should be constrained using specific Michaelis-Menten kinetic equations for each carbon source. Contrarily, ecYeast8 simulations show a good agreement with experimental data as the order of substrate consumption in this model is determined by the relative protein cost for substrate consumption as well as the biomass yield on the different carbon sources (Table 5.1, Figure 5.3D-F).

Table 5.1: Relative protein cost for consumption of different substrates and biomass yield per C-mol. The relative protein cost is calculated as the flux through the protein pool reaction required to consume 1 mmol of substrate divided by the same flux required for the consumption of 1 mmol of glucose.

Substrate	Relative protein cost	Biomass yield (g_{DW}/C-mol)
Glucose	1.00	0.43
Fructose	1.25	0.30
Mannose	1.27	0.30
Sucrose	2.18	0.20

When sucrose and glucose are the substrates, the model predicts three phases characterized by the use of different carbon sources. First, all the available free glucose is consumed, as it is the substrate with the lowest protein cost (Table 5.1). In the second phase, sucrose is hydrolyzed, sucrose-derived glucose is consumed and fructose accumulates. The highest protein cost of sucrose is caused by the simultaneous consumption of glucose and fructose. However, during dFBA simulation the accumulation of glucose and fructose in the reactor is allowed and the only additional cost of sucrose consumption is caused by the need to hydrolyze the disaccharide by the invertase enzyme. After hydrolysis, ecYeast8 predicts glucose consumption and fructose accumulation due to the lower protein cost of glucose degradation (Table 5.1). The third phase is characterized by fructose consumption, with a higher protein cost compared to glucose use caused by a higher flux through the glucose-6-phosphate (G6P) isomerase. According to ecYeast8, this enzyme converts G6P to fructose-6-phosphate (F6P) during growth on glucose and catalyzes the reversible reaction during fructose growth with a higher flux. The fact that these three phases

are also observed experimentally suggests that carbon catabolite repression (CCR) of sucrose, fructose and ethanol exerted by glucose is essential to achieve maximum growth rate when considering the limitation of protein content in the cells (Figure 5.3D).

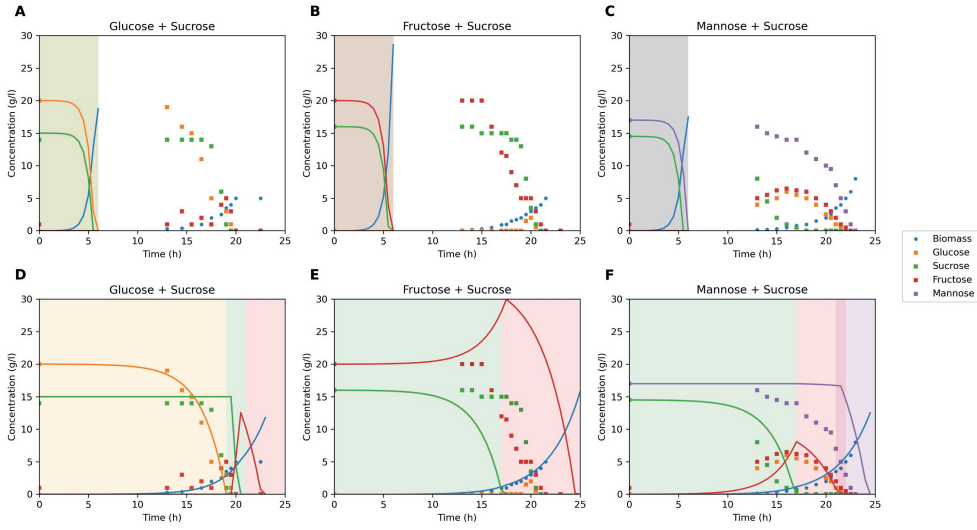


Figure 5.3: Batch simulations of *S. cerevisiae* DGI342 with two carbon sources using Yeast 8 (**A-C**) and ecYeast8 (**D-F**) compared to experimental data (symbols) [261]. Colored areas represent different substrate consumption phases predicted by the model: glucose consumption (orange), sucrose hydrolysis and glucose consumption (green), fructose consumption (red), and mannose consumption (purple).

When the carbon sources are sucrose and fructose, the model repeats phases two and three. First, sucrose is hydrolyzed, the sucrose-derived glucose is consumed and fructose accumulates. Then, only after glucose depletion, the model consumes the available fructose (Figure 5.3E).

Simulations with sucrose and mannose show similar results. First, the model predicts the consumption of sucrose-derived glucose and fructose accumulation. Fructose consumption only starts after glucose depletion. Mannose consumption starts at a low rate at the end of the fructose consumption phase and continues then at a higher rate due to the higher protein cost required for its degradation. This cost is caused by the need to convert mannose to F6P, reactions catalyzed by mannokinase and mannose-6-phosphate isomerase (Figure 5.3F).

Although the model does not predict the initial consumption of fructose in simulations with fructose and sucrose, or initial glucose accumulation and simultaneous consumption of glucose, fructose, and mannose in sucrose and mannose simulations, the protein availability constraint is enough to accurately predict sucrose hydrolysis as well as fructose and mannose consumption rates. Besides, the combination of ecYeast8 and dFBA improved predictions by explicitly modeling the inhibitory effect of glucose, fructose, sucrose, and mannose on the uptake rates of the other carbon sources (Sup. Figure 5.2).

Simulation of a Δ*pdc* lactate producing *S. cerevisiae* strain

Yeast8 and ecYeast8 were modified to simulate a *S. cerevisiae* strain without pyruvate decarboxylase activity, laboratory evolved to tolerate high glucose concentrations and engineered to produce lactate [253]. dFBA simulations with Yeast8 and ecYeast8 were in agreement with experimental data (Figure 5.4A). According to Van Maris et al., during the first 24 h of the fermentation oxygen was supplied in excess to the reactor and cells were only limited by glucose availability. After 24 h cells suffered oxygen limitation, which was simulated constraining the oxygen uptake reaction. The oxygen limitation continued after 75 h when an additional 100 g pulse of glucose was added to the reactor [253].

During laboratory evolution fastest growers were selected, obtaining a final strain with a maximum growth rate of 0.13 h^{-1} [253]. Although the concept of laboratory evolution is in agreement with the use of biomass growth as an objective function during FBA, Yeast8, and ecYeast8 predicted higher maximum growth rates. Therefore, the upper bound of the biomass reactions had to be constrained to match the experimental value.

During the glucose limitation phase both models, Yeast8 and ecYeast8, were limited by glucose availability determined by a Michaelis-Menten equation. Besides ecYeast8 was limited by the protein pool constraint which resulted in the prediction of a lower glucose uptake rate by this model (Figure 5.4B). In this period oxygen uptake rates predicted by Yeast8 were unreasonably high and lactate production was not predicted (Figure 5.4B,C). Contrarily, the limitation in protein availability of ecYeast8 resulted in realistic predictions of oxygen uptake and lactate production rates. After 24 h the limitation in oxygen uptake resulted in a 99.88% decrease in oxygen uptake by Yeast8 (from 34 mmol/g_{DW}/h to 0.04 mmol/g_{DW}/h) and 93% decrease in ecYeast8 (from 0.58 to 0.04 mmol/g_{DW}/h) (Figure 5.4). After the introduction of this limitation, there were no significant differences in flux predictions by both models.

Besides lactate production, simulations by ecYeast8 resulted in succinate and glycerol production at concentrations similar to experimental measurements [253]. Simulations with Yeast8 only resulted in glycerol and succinate production once oxygen uptake was limited and additional by-products such as citrate or arginine were exported by the model.

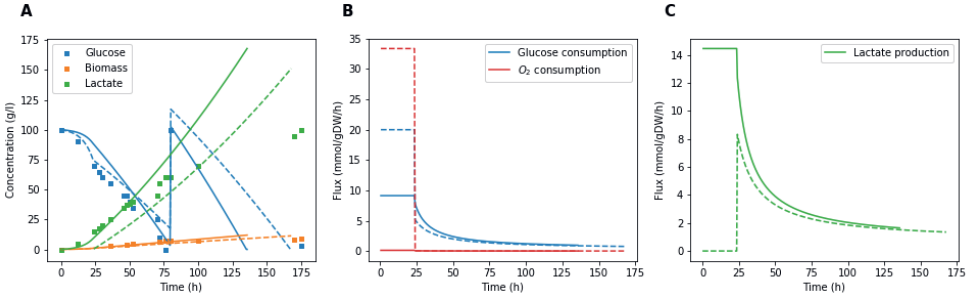


Figure 5.4: Batch growth simulation of a Δ*pdc*, lactate producing *S. cerevisiae* strain using EcYeast8 (-) and Yeast8 (- -) compared to experimental data (□).

Although both models performed well when additional constraints were applied, the protein availability constraint was enough to predict lactate production in oxygen excess conditions suggesting the potential of combining enzyme-constrained models and dFBA for cell factory simulations (Figure 5.4). The disagreement between model predictions and experimental data observed during the second half of the simulations is probably caused by growth inhibition due to product toxicity. The dFBA framework would allow to include this inhibition linking the upper bound of the biomass reaction to the reactor concentration of the toxic compound.

Flux sampling as a tool to explore metabolism at different growth rates

When modeling cell metabolism, FBA only provides one of the multiple flux distributions that result in the optimization of the chosen objective function. Flux sampling algorithms solve this problem by providing possible flux distributions of metabolic reactions that satisfy mass balance constraints [64]. Due to the better performance of ecYeast8 when simulating the consumption and production of metabolites in different reactor settings, we tested how flux sampling can be applied to study intracellular fluxes.

During simulations with ecYeast8, all the glyceraldehyde-3-phosphate was produced through the pentose phosphate pathway. To ensure experimentally observed flux through phospho-fructo kinase and fructose bisphosphate aldolase, the reversible transaldolase reaction was blocked before sampling [262]. Also, the reduction of the tricarboxylic acid cycle (TCA) intermediates in the cytoplasm was avoided to favor the production of NADH in the cytoplasm [263]. Last, the model was re-scaled to avoid stoichiometric coefficients below solver tolerance that caused numerical instability. For each simulation, the obtained flux distributions represent the metabolism of *S. cerevisiae* cells growing in a chemostat with a specific dilution rate at steady state. The simulation at maximum growth rate represents the metabolism of cells growing exponentially in a batch reactor. Sampling results can be found at Gitlab.

In general, we observed good agreement between predicted fluxes and experimental measurements. For example, the flux through TCA reactions decreases with growth rate and, at the maximum growth rate, sampling results show the operation of the TCA cycle as two different branches (zero flux through α-ketoglutarate dehydrogenase (KGD), succinyl-CoA synthetase (SCL) and fumarase (FUM)) [262, 264, 265]. At these high growth rates, relative flux to the pentose phosphate pathway (PPP) decreases and is directed towards glycolysis and ethanol formation (Figure 5.5A). As expected, the variability of the fluxes decreases at increasing growth rates as a result of a more limited solution space. At higher growth rates protein availability becomes limiting and alternative pathways are no longer feasible.

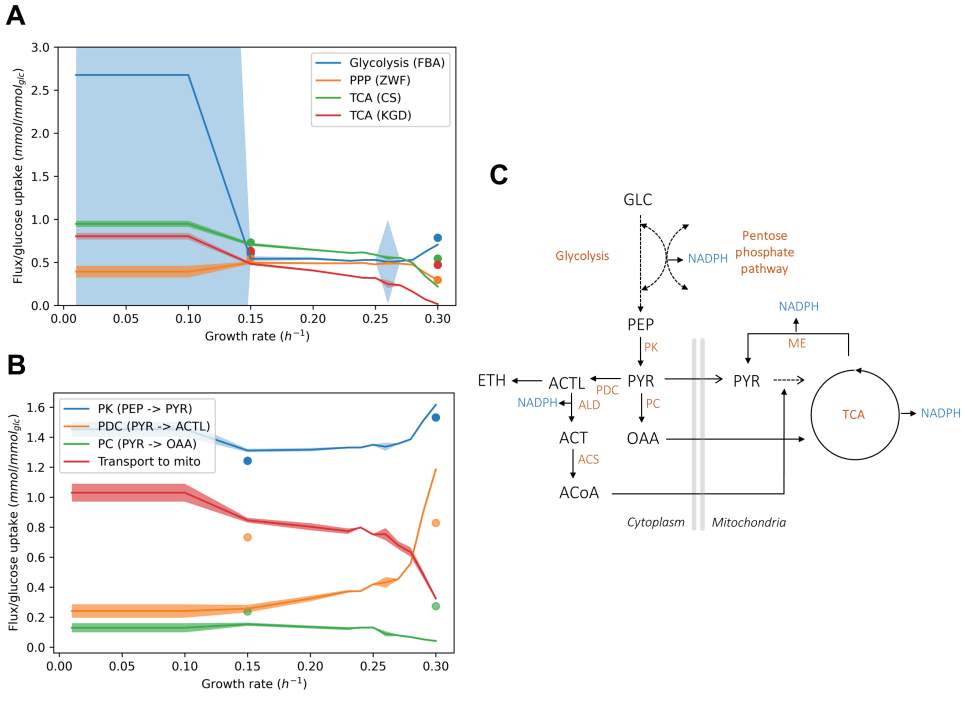


Figure 5.5: Comparison of flux sampling results with ecYeast8 (median ± MAD) and 13C flux analysis data (symbols) [262]. **A.** Fluxes relative to the glucose uptake at different growth rates of the glycolytic enzyme fructose bis-phosphate aldolase (FBA), the PPP enzyme glucose-6-phosphate dehydrogenase (ZWF), the tricarboxilic acid cycle enzymes citrate synthase (CS) and α-ketoglutarate dehydrogenase (α-KGD). **B.** Fluxes relative to the glucose uptake of enzymes involved in the pyruvate node (pyruvate kinase, PK; pyruvate decarboxylse, PDC; pyruvate carboxylase, PC) and relative transport flux of pyruvate to mitochondria (mito). **C.** Schematic representation of the pyruvate node (GLC, glucose; PEP, phosphoenol pyruvate; PYR, pyruvate; OAA, oxaloacetate; ACTL, acetaldehyde; ETH, ethanol; ACT, acetate; ACoA, acetyl coenzyme A; ALD, acetaldehyde dehydrogenase; ACS, acetyl CoA synthase; ME, malic enzyme). In A and B lines represent the median flux value obtained from 10,000 samples and shaded areas represent the median absolute deviation.

As a test case on the use of flux sampling to study metabolism, we focused on the predicted flux distributions in the pyruvate node and compared them to experimental data (Figure 5.5B, C) [262]. Pyruvate kinase (PK) is the main source of cytoplasmic pyruvate and, in agreement with literature, the model predicts constant relative flux at growth rates below the critical dilution rate and increasing relative flux at higher growth rates. Whilst flux predictions of PK and pyruvate decarboxylase (PDC) follow the same trend as experimental data, pyruvate carboxylase (PC) shows the opposite behavior (Figure 5.5B). Frick and Wittmann propose that at high growth rates, pyruvate conversion to acetyl-CoA (by PDC, ALD, and ACS) and subsequent transport to mitochondria saturates. The extra pyruvate is then converted to oxaloacetate by PC, which is transported to the mitochondria and converted back to pyruvate by the malic enzyme (ME). In this way, the mitochondrial pyruvate pool, required for acetyl-CoA and amino acid synthesis, is replenished [262,

266]. However, the model predicts a relative flux increase through PDC, no saturation in the cytoplasmic conversion of pyruvate to acetyl-CoA, and, as a result, fails to predict the experimentally observed flux increase through PC (Figure 5.5C). Although model predictions show a decrease in relative pyruvate transport to the mitochondria, transport is enough to cover mitochondrial pyruvate requirements and the experimentally observed flux increase through ME is not predicted by the model (Figure 5.5C). In fact, free movement of metabolites across compartments is allowed in ecYeast8 as transporters are not part of the protein pool. Therefore, inaccurate flux predictions are expected when the transport of metabolites across compartments is the limiting factor.

Discussion

EcModels add an additional layer of information to traditional GEMs based on the limited capacity of the cells to synthesize proteins, which results in more accurate predictions of extracellular fluxes during chemostat, batch, and fed-batch growth of different *S. cerevisiae* strains. In chemostat simulations, ecYeast8 corrects the inability of Yeast8 to predict the critical dilution rate and subsequent decrease in biomass concentration and ethanol production. Similarly, during batch simulations, ecYeast8 corrects the inability of Yeast8 to predict the Crabtree effect as well as the order and rate of consumption of several carbon sources.

De Groot et al. show that GEMs predict overflow metabolism when two growth-limiting constraints are hit regardless of their biological interpretation [267]. Therefore, Yeast8 can be modified to predict overflow metabolism by adding a second constraint such as a maximum oxygen uptake rate [268]. However, ecYeast8 not only predicts respiro-fermentative metabolism at growth rates higher than the critical dilution rate but, when combined with dFBA, it also accurately describes ethanol production and consumption during exponential growth, the preferred consumption order of different carbon sources as well as product production rates (Figures 5.2, 5.3, 5.4). In traditional GEMs, the flux through reactions required for growth is not constrained, so the model adjusts these fluxes to obtain the desired growth rate, which results in an inaccurate description of metabolism. EcYeast8 breaks the linear dependency between fluxes and growth rate and shows accurate intracellular flux predictions (Figure 5.5). Simulating this behavior with Yeast8 is only possible upon an iterative, case-dependent design of condition-specific constraints [269].

The parameter with the largest influence on the simulations is the upper bound of the protein exchange reaction, which represents enzyme availability. This parameter determines the maximum growth rate in batch reactors, the critical dilution rate in continuous cultures, and the uptake rates of substrates. In the absence of proteomic data Sánchez et al. assume constant protein availability for a given strain and process and provide two different values depending on the simulation of chemostat or batch growth [67]. We showed that increasing this parameter was required to simulate batch growth on different carbon sources and that it should be adjusted to accurately simulate chemostat growth of strains with different protein content [258]. Interestingly, the effect of the protein availability constraint implicitly accounts for protein synthesis costs reducing the

impact of the maintenance requirements during the simulations. Therefore, the constraint in protein availability can be understood in terms of the limited space in the cell, but also in terms of limited energy available for protein synthesis. Besides, simulations with different carbon sources suggested that the order of substrate use can be partially explained by the associated protein cost required for its consumption. When considering the limitation of protein content in the cells, CCR is essential to achieve the maximum growth rate.

dFBA is a valuable tool to predict the dynamic behavior of engineered strains in a bioreactor [249, 250, 251]. Whilst FBA allows the comparison of yields between engineered strains, dFBA simulates dynamic processes allowing the comparison of final titters and productivities which depend on the strain and the bio-process. We showed here how the combination of ecYeast8 with dFBA improved predicted metabolic changes in response to the operation of a reactor without additional constraints. We used simulations on a mixture of carbon sources to show how predictions can be further improved by incorporating regulation-related constraints to the dFBA framework (Sup. Figure 5.2). We showed the potential of this method to aid the design of bio-processes including the prediction of the metabolism of engineered cells in a reactor and changes in cell metabolism due to changes in operational conditions such as co-feeds. This framework can be extended to include other important process parameters such as temperature [187].

The accurate prediction of intracellular metabolic fluxes is a desired feature for models aiming to find and compare metabolic engineering strategies to improve the production of a target metabolite. To the best of our knowledge, this study is the first report on how to combine flux sampling and ecModels to study intracellular flux predictions, avoiding the necessity to fix an objective function and allowing the coverage of the whole solution space [64]. While previous studies focused on the prediction of intracellular fluxes at the maximum growth rate, we have compared flux predictions covering *S. cerevisiae* full range of growth rates [263]. Despite the substantially improved predictive power of the model, the protein availability constraint was not enough to yield accurate predictions of all intracellular fluxes due to the highly dimensional solution space and the absence of regulatory information in the model (Figure 5.5B). Using proteomic data instead of a single constraint on the protein content of the cells, considering space limitation in cell membranes or the creation of ensemble models is expected to further improve flux predictions when these models are applied to strain design, reducing the prediction of incorrect knock-out and overexpression targets [67, 270, 271].

In conclusion, we introduced flux sampling as a tool to analyze intracellular flux predictions of ecModels, of major importance for model-guided strain design. As parameters in the reactor, as well as genetic modifications, affect flux predictions, the successful combination of ecModels and dFBA allows the comparison of yields and productivities among different strains and (dynamic) production processes. This model and simulation framework therefore provides the means for more accurate and realistic designs of cell-based processes increasing their usefulness for industrial applications.

Declaration of interest

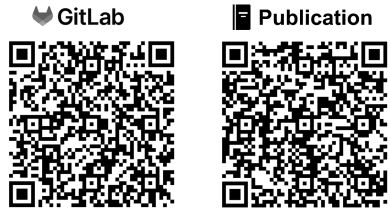
Joep Schmitz is employed by DSM and Vitor A. P. Martins dos Santos has interests in LifeGlimmer GmbH.

Acknowledgment

This project was founded by NWO (project number GSGT.2019.008) and IBISBA (H2020 project numbers 730976 and 871118).

Data availability

Scripts and data are available at Gitlab and the published version of the article.



Supplementary methods

EcYeast8 model modifications

Model re-scaling

In ecModels enzymes are treated as metabolites which stoichiometric coefficient is $1/k_{cat}$. In ecYeast8 k_{cat} values expand 10 orders of magnitude (1 to 10^{10} h^{-1}) resulting in stoichiometric coefficients below solver tolerance (10^{-6}) and numerical instability of the model during flux sampling. To minimize the problem the range of k_{cat} values was reduced so the maximum allowed k_{cat} value was 10^6. Besides, all $1/k_{cat}$ coefficients and the protein pool exchange upper bound were scaled by 10^3 to reduce the impact of rounding errors on flux predictions through enzyme usage reactions. These modifications did not change flux predictions as the contribution of extremely efficient enzymes (*i.e.* $k_{cat} > 10^6$) to the protein pool is negligible.

Model constraints

For all the simulations the upper bounds of exchange reactions to produce acetaldehyde, 2,3-butanediol, glycine, acetate, and pyruvate were constrained to match experimental measurements [67]. Also, the transport of serine from the mitochondria to the cytoplasm (r_2045_REV) and the cytoplasmic $NADP^+$ dependent conversion of isocitrate to 2-oxoglutarate (r_0659No1) were blocked as described in Sánchez et al. [67]. Besides, the reversible transaldolase reaction (r_1048_REVNo1), the reversible cytoplamic reaction of malate dehydrogenase (r_0713_REVNo1), the fumarate reductase reaction in the cytoplasm (r_1000No1), the reversible isocitrate dehydrogenase reaction in the cytoplasm (r_0659_REVNo1) and the glutamate decarboxylase reaction (r_0469No1) were blocked by constraining to zero their upper and lower bounds.

Experimental data

Experimental data of *S. cerevisiae* growth in chemostat and batch reactors was obtained from literature (Sup. Table 5.1). Fed-batch cultures of *S. cerevisiae* CEN.PK-113-7D were performed in a 1 l working volume of a stirred fermenter (DASGIP parallel bioreactor system, Eppendorf). Throughout the fermentations the pH was kept at 5.1 and the temperature was kept at 30°C. The batch medium (400 g) contained: 2.5 g/kg glucose, 1g/kg $(NH)_4SO_4$, 10 g/kg KH_2PO_4, 4 g/kg $MgSO_4$·7H2O, 0.3 g/kg $CaCl_2$·$2H_2O$ and vitamins and trace elements according to Verduyn et al. [272]. After 4 hours of growth on the batch medium, the aerobic fed-batch phase was started with an exponential feed profile supporting a growth rate of 0.05 h^{-1}. The composition of the feed medium was 209 g/kg glucose, 7.67 g/kg ethanol, 2g/kg $(NH)_4SO_4$, 20 g/kg KH_2PO_4, 8g/kg $MgSO_4$·$7H_2O$, 0.6 g/kg $CaCl_2$·$2H_2O$ and vitamins and trace elements according to Verduyn et al. [272]. Samples for biomass concentration determination were obtained every 24 hours and actual oxygen uptake rate and CO_2 production rate were determined using off-gas analysis.

Sup. Table 5.1: Summary of experimental data used in this study. *CEN.PK 113.7D pdc1(-6.-2)::loxP pdc5(-6.-2)::loxP pdc6(-6.-2)::lox P ura3-52 YEpLplLDH

Reactor operation	Carbon source	Strain	Reference
Chemostat	Glucose	CBS8066	[254]
Chemostat	Glucose	DS28911	[255]
Chemostat	Glucose	H1022	[256]
Chemostat	Glucose	CEN.PK 113.7D	[259]
Batch	Glucose	H1022	[260]
Fed-batch	Glucose	CEN.PK 113.7D	This study
Batch	Sucrose + Glucose	DGI342	[261]
Batch	Sucrose + Fructose	DGI342	[261]
Batch	Sucrose + Mannose	DGI342	[261]
Batch	Glucose	GCSI-L*	[253]

Chemostat, batch, and fed-batch simulations

Chemostat reactors are in steady state and have a constant inflow and outflow of media at a rate F (l/h). Note that, the dilution rate, D, (h^{-1}) is defined as F /V where V represents the reactor volume (l). The biomass balance shows that D equals the cell growth rate (μ, h^{-1}). Given D and the glucose concentration in the feed ($c_{s,in}$, mmol/l), FBA is used to calculate the glucose consumption rate (q_s, mmol/g_{DW}/h), and the glucose mass balance is used to calculate the cell concentration (c_x, g_{DW}/l). Similarly, the concentration of product i (c_i , mmol/l) is calculated with the product mass balance using its production rate (q_i, mmol/g_{DW}/h) obtained by FBA (Sup. Table 5.2).

In batch reactors and during the batch phase of fed-batch reactors biomass and products accumulate, substrates are depleted at a rate determined by the Michaelis-Menten equation, and the glucose mass balance is used to calculate the remaining glucose in the reactor (M_s, mmol). During the feeding phase of fed-batch reactors, the reactors are fed with media containing substrates at a rate F (l/h), there is no accumulation of glucose (*i.e.* $dM_s/dt = 0$) and the glucose mass balance is used to constrain the glucose uptake rate (q_s, mmol/g_{DW}/h). FBA with q_s as constraint is used to calculate the growth rate (μ, h^{-1}) and the production rate of other metabolites (q_i, mmol/g_{DW}/h). Mass balances are used to calculate the biomass mass in the reactor (M_x, g_{DW}) as well as the product mass (M_i, mmol). Note that mass is used instead of concentrations as the volume in the reactor changes during the feeding phase, concentrations are calculated as $M/V(t)$, where V(t) is the liquid volume in the reactor at time (t) (Sup. Table 5.2).

Sup. Table 5.2: Glucose, biomass and product mass balances in chemostat, batch and fed-batch reactors. Chemostat reactors are in steady state and have a constant inflow and outflow of media. In batch reactors biomass and products accumulate and substrates are depleted. During the feeding phase of fed-batch reactors, the reactors are fed with media containing substrates. D, dilution rate (h^{-1}); $M_{s,in}$, glucose mass in the feed (mmol); q_s, glucose uptake rate (mmol/g_{DW}/h); M_x, biomass mass in the reactor (g_{DW}); M_s, glucose mass in the reactor (mmol); μ, growth rate (h^{-1}); q_i, production rate of product i (mmol/g_{DW}/h); M_i, mass of product i in the reactor (mmol), t time (h); F, feed rate (l/h); $c_{s,in}$, glucose concentration in the feed (mmol/l).

Chemostat

Glucose	$0 = D * M_{s,in} - q_s * M_x - D * M_s$
Biomass	$0 = \mu * M_x - D * M_x$
Product i	$0 = q_i * M_x - D * M_i$

Batch and fed-batch

Glucose	Batch: $\frac{dM_s}{dt} = F * c_{s,in} - q_s * M_x(t)$
	Feed: $0 = F * c_{s,in} - q_s * M_x$
Biomass	$\frac{dM_x}{dt} = \mu * M_x(t)$
Product i	$\frac{dM_i}{dt} = q_i(t) * M_x(t)$

Simulation of mixed carbon fermentations including additional regulation with ecYeast8

Consumption of combinations of sucrose with glucose, fructose, and mannose was simulated using ecYeast8 and including additional regulatory constraints in the dFBA framework.

First, repression of fructose consumption by glucose was simulated by constraining the upper bound of the fructose uptake reaction (r_1709_REV) to zero when glucose concentration in the media was higher than 1 mmol/l. Similarly, repression of glucose consumption by fructose was simulated constraining the glucose uptake and transport reactions (r_1714 REV and r_1166) to zero when fructose concentration in the media was above 1 mmol/l. When sucrose and mannose were the initial carbon sources, the experimentally observed delay between sucrose hydrolysis and glucose and fructose consumption was simulated constraining their transport reactions (r_1166 and r_1134) to zero when sucrose was present in the media.

Second, sucrose hydrolysis was forced when glucose and fructose concentrations in the media were below 15 g/l by setting a lower bound to reaction r_2058_REV. If glucose and fructose were initially present in the media the bound was 3.5 mmol/g_{DW}/h and 20 mmol/g_{DW}/h otherwise.

In all cases, lower and upper bounds of glucose, fructose, and mannose uptake as well as sucrose hydrolysis were set to zero when the carbon sources were not present in the media.

Last, during simulations with fructose and mannose as initial carbon sources, the protein exchange upper bound was increased by 75% compared to batch simulations with glucose and the secretion reactions for ethanol, acetate, pyruvate, acetaldehyde, and 2,3-butanediol were unconstrained.

Simulation of a Δ*pdc* lactate producing *S. cerevisiae* strain

Yeast8 and ecYeast8 were modified to simulate a strain without a pyruvate decarboxylase (PDC) activity and expressing the lactate dehydrogenase gene from *Lactobacillus plantarum* [253]. To simulate the knockout, bounds of reactions r_0959 and r_0960 (Yeast8) and arm r_0959 and arm_r_0960 (ecYeast8) were set to zero. The lactate dehydrogenase reaction was added to Yeast8 (Equation 5.1). Three reactions were added to ecYeast8: the forward and reverse reactions of the lactate dehydrogenase (Equations 5.2, 5.3) and a draw reaction for the LDH protein (Equation 5.4). Values of k_{cat} (40 s^{-1}) and molecular weight (39 kDa) were obtained from BRENDA [47].

$$Lactate + NAD^+ \leftrightarrow Pyruvate + NADH + H^+ \tag{5.1}$$

$$Lactate + NAD^+ + \frac{1}{k_{cat}} \cdot LDH \rightarrow Pyruvate + NADH + H^+ \tag{5.2}$$

$$Pyruvate + NADH + H^+ + \frac{1}{k_{cat}} \cdot LDH \rightarrow Lactate + NAD^+ \tag{5.3}$$

$$MW \cdot prot_pool \rightarrow LDH \tag{5.4}$$

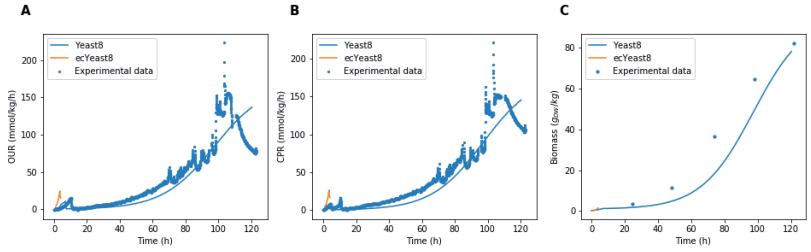
Batch growth of both models was simulated using dFBA. The growth reaction (r_2111) upper bound was constrained to 0.13 h^{-1} to simulate the maximum growth rate observed experimentally [253]. Cells were simulated in a 1 l reactor operated as a batch with 100 g/l of initial glucose. Eighty hours after inoculation 100 g of glucose were added to the reactor and this pulse was included in the simulations. According to Van Maris et al., cells experienced oxygen limitation from 24 h after inoculation until the end of the process [253]. Oxygen limitation was simulated constraining the oxygen uptake reaction (r_1992 in Yeast8 and r_1992_REV in ecYeast8) to q_o assuming pseudo-steady state for oxygen:

$$0 = max\_O_2\_transfer - q_o \cdot c_x \tag{5.5}$$

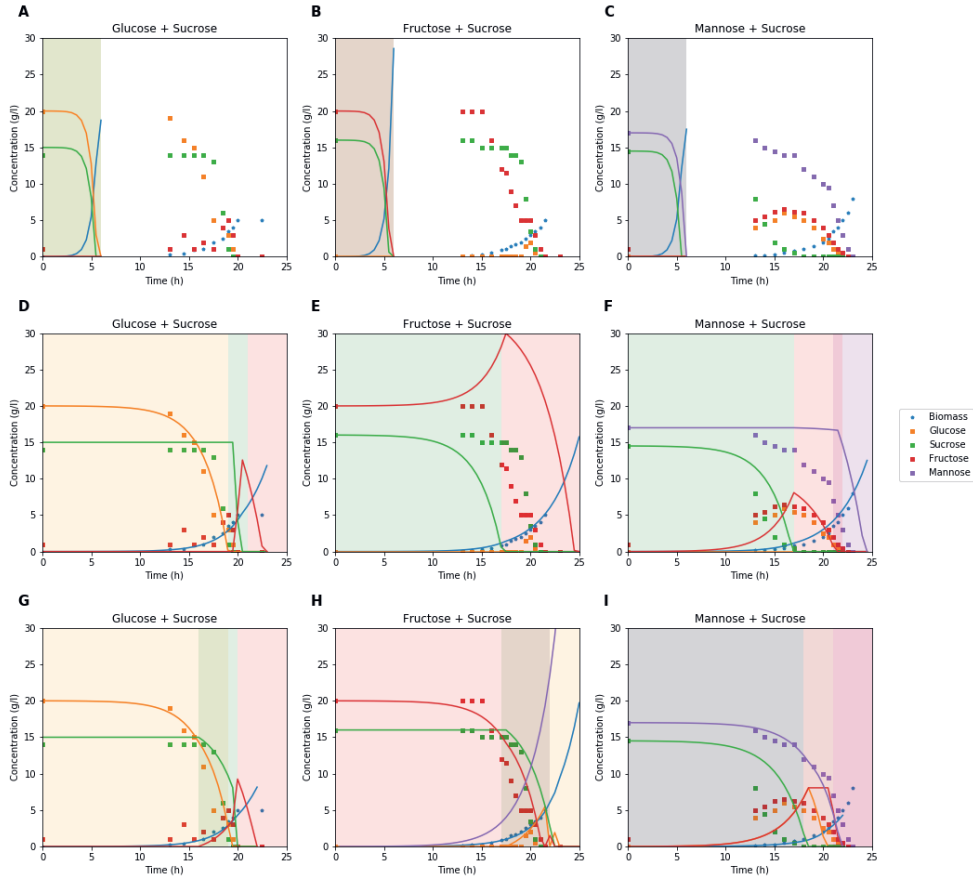
where the maximum O_2 transfer to the reactor ($max\_O_2\_transfer$) was calculated based on experimental data and c_x is the predicted biomass concentration in the previous time step [253].

In agreement with experimental data, during ecYeas8 simulations export of products different than biomass, lactate, succinate, and glycerol was avoided constraining their production reactions [253]. The same approach resulted in infeasible solutions with Yeast8 and therefore production of alternative products was allowed.

Supplementary Figures



Sup. Figure 5.1: Simulations of OUR (**A**), CPR (**B**), and biomass concentration (**C**) of *S. cerevisiae* CEN.PK 113.7D grown with an exponential glucose feed. Peaks in experimental OUR and CPR data were caused by sampling of the reactor.



Sup. Figure 5.2: Batch simulations of *S. cerevisiae* DGI342 with two carbon sources using Yeast8 (**A-C**), ecYeast8 (**D-F**) and ecYeast8 with additional regulation (**G-H**) compared to experimental data (symbols) [261]. Colored areas represent different phases in the process: glucose consumption (orange), sucrose hydrolysis and glucose consumption (green), fructose consumption (red), and mannose consumption (purple).

CHAPTER 6

Responses of *Pseudomonas putida* to glucose and oxygen limitations

Sara Moreno Paz, Cristina Furlan, Vitor A. P. Martins dos Santos, Christos Batianis#, María Suárez Diez#

#Jointly supervised this work

Abstract

Understanding cell physiology during oxygen-limited growth is essential, as reducing the oxygen demand of microbial bio-conversions can facilitate their implementation. Similarly, glucose limitation is often employed in industrial bioprocesses. Therefore, we investigated the adaptive responses of the obligate aerobic bacteria *Pseudomonas putida* KT2440 to varying oxygen levels in chemostat bioreactors, and compared glucose-limited and oxygen-limited cultures. Under oxygen-limited growth, *P. putida* exhibited a decrease in cell concentration, coupled with an up to 59% increase in biomass yield and a reduced glucose uptake. Notably, the increased yield was caused by the lack of pyoverdine production in this condition, compared to the observed production during slow, glucose-limited growth. Transcriptomic and proteomic samples were analyzed considering the aging of the culture to identify changes specifically related to varying nutrient limitations. While 923 differentially expressed genes, specific to oxygen-limited growth were identified, only seven differentially abundant proteins were found. Genes up-regulated during oxygen limitation were associated with respiratory chain functions, while down-regulated genes were related to a few catabolic processes including β-oxidation and protein maturation. After eight days of oxygen-limited growth, excess oxygen was reintroduced, and the cells exhibited substantial recovery, reaching cell concentrations and glucose uptake levels comparable to those observed during the initial glucose-limited growth. Overall, our findings suggest *P. putida*'s resilience to long-term oxygen-limited growth, which can be applied to reduce energy requirements in industrial-scale bioprocesses and benefit the production of reduced products. Moreover, we unveil the carbon loss of slow-growing, glucose-limited cells caused by pyoverdine production.

Introduction

Pseudomonas putida KT2440 has gained recognition as a versatile chassis for various metabolic engineering applications. Its unique attributes, including the ability to adapt to high organic solvent concentrations and oxidative stress, coupled with its straightforward nutritional requirements, rapid growth, and versatile metabolism, position it as an ideal candidate for industrial bio-transformations [105, 121, 122, 273, 274]. However, its obligate aerobic nature hinders production in large-scale bioreactors where oxygen-limited zones are inevitable [275]. While oxygen limitation could affect the organism's growth and performance, maintaining high oxygen tensions increases the energy and cooling requirements of the bioprocess [275, 276]. Agitation alone is estimated to require approximately 1 kW of power for every cubic meter of bioreactor volume and the sum of agitation and aeration is estimated to require around 3 kW/m^3 [277]. Although anoxic regimes are not always necessary, the development of low-oxygen production processes would have a profound impact on lowering costs during scale-up and, consequently, on the adoption of microbial bio-conversions [277].

The physiology of *P. putida* has been studied in the presence of different carbon sources [278], toxic compounds [279, 280], and limited nutrients such as nitrogen, phosphorus or iron [274, 281, 282]. The ability of *P. putida* to endure temporal glucose limitation and starvation through internal 3-hydroxy alkanoates (3-HA) utilization, a precursor to polyhydroxyalkanoates (PHA), and amino acid catabolism is well-documented [283]. Similarly, the physiology of *P. putida* grown in glucose-limited chemostat reactors has been studied at different growth rates [284]. However, its physiological and genetic responses to oxygen limitation remain insufficiently explored. This knowledge gap is significant, as understanding these responses is essential for effectively operating and optimizing this bacteria in large-scale industrial environments.

Several studies have attempted to engineer anaerobic *P. putida* cells by modifying ATP generation pathways, employing redox mediators in bioelectrochemical systems, or altering essential oxygen-dependent reactions [285, 286, 287, 288, 289, 290]. Despite these efforts, successful growth under anoxic conditions has not been achieved, with only Kampers et al. reporting improved growth under microoxic conditions, highlighting the complexity of *P. putida*'s metabolic needs [289]. To our knowledge, the only study that investigated the response of *P. putida* to oxygen limitation using omics technologies is the work of Demling et al. [291]. They explored its adaptability to oxygen gradients mimicking industrial reactors, providing insights into its physiological and proteomic adjustments under these changing conditions. They revealed that, while fluctuating oxygen levels led to a deceleration in growth, the final biomass and product concentrations remained unaffected. Moreover, the minimal proteomic changes observed in cells exposed to oscillating oxygen availability underscored *P. putida*'s capacity to cope with such environmental variations [291].

While Demling et al. showed the ability of *P. putida* to rapidly recover from oxygen starvation when resupplied with oxygen [291], Ankenbauer et al., and Mutyala et al. proposed microaerobic growth of this microorganism to increase production of isobutanol and succinate, respectively [292,

293]. In this way, the growth of *P. putida* under microoxic conditions is not only important to endure gradients in industrial fermentors, but can also be used as a strategy to improve production.

Here, we study the physiological, transcriptomic, and proteomic responses of *P. putida* KT2440 cells grown in glucose- and oxygen-limited chemostats. This setup allowed us to study the cellular adaptations to long-term oxygen limitation, a scenario scarcely explored, yet crucial for understanding *P. putida*'s performance in industrial settings. Although the intracellular mechanisms that cope with this prolonged limitation remain ambiguous, we show the resilience of *P. putida* to oxygen-limited conditions. Notably, the use of glucose-limited cells as a reference condition revealed a significant reduction of biomass yield during this limitation caused by the production of pyoverdine despite the use of iron-rich media. This decreased yield is corrected upon the enforcement of oxygen-limited growth. This study, therefore, provides a comprehensive view of *P. putida*'s adaptive capabilities, laying a foundation for its efficient utilization in industrial bioprocesses.

Materials and methods

Strains and media

P. putida KT2440 was used in all the experiments. Luria-Bertani (LB) medium containing 5 g/l yeast extract, 10 g/l tryptone and 10 g/l NaCl was used for overnight pre-culture cultivations. M9 minimal medium (6 g/l Na_2HPO_4, 3 g/l KH_2PO_4, 1.4 g/l $(NH_4)_2SO_4$, 0.5 g/l NaCl, 0.2 g/l $MgSO_4 \cdot 7H_2O$ and 2.5 ml/l trace elements) was used for bioreactor cultivations [294]. Media was supplemented with 80 mM of glucose and 0.6 ml/l of antifoam 204 (Sigma-Aldrich).

Bioreactor cultivation

Cells from the pre-cultures were washed with minimal media and bioreactors were inoculated with an initial OD600 of 0.3. The cultivation was performed in 1 l bioreactors connected to a Biostat Q plus tower and controlled through a Biopat DCU tower via MFCS win 3.0 (Sartorius-Stedim, Gottingen, Germany). Bioreactors containing 500 ml of media were maintained at 30°C and pH was controlled at 7 using 15% w/v NH_4OH. During batch fermentations, the dissolved oxygen (DO) was maintained above 30% by increasing the stirrer speed from 400 rpm up to 1200 rpm with an airflow rate of 0.6 l/min. Exhaust gas composition was monitored online by BlueSense CO_2 and O_2 infrared gas analyzers (Herten, Germany). The batch phase was continued by a chemostat phase upon glucose depletion. A dilution rate of 0.1 h^{-1} was controlled by peristaltic pumps that fed media at a rate of 0.83 ml/min. The exhaust media was controlled by a sensor that ensured constant volume in the fermentors. After five volume changes of the reactor media, steady state was determined when, during three consecutive samples, taken 24 h apart, values for OD600 in the effluent and/or the reactor, metabolite concentrations, oxygen, and CO_2 concentration in the exhaust gas, and DO varied less than 10%. Liquid and gas flow rates were monitored and adjusted when required.

During the chemostat cultivation, four sequential steady states were achieved (Figure 6.1A). The first steady state was characterized by glucose limitation and the DO was maintained above 30% adjusting the stirrer speed and using an air flow rate of 0.6 l/min. After the steady state was reached, the stirrer speed was maintained constant at 750 rpm and the gas phase was changed to a mixture of 95% N_2 and 5% O_2. After a new steady state, the concentration of O_2 was decreased to 2% of the gas. Finally, the gas inflow was switched back to air and the DO was maintained above 30%. Samples for RNA-seq analysis, proteomics, HPLC, and cell dry weight (CDW) determination were collected at every steady state.

Analytical methods

Cell density was determined either at 600 nm (OD600) using Diluphotometer IMPLEN or measuring cell dry weight (CDW) (in g_{DW}/l) at a given time point. CDW was determined by filtering 5 ml fermentation broth using pre-weighted dry microfilters (0.4 µm pore size). After drying for 24 h at 105°C, the filters were cooled down and weighed to calculate the cell dry mass.

Concentrations of glucose, gluconic acid, ∝-keto gluconic acid, and other organic acids in the culture supernatant were determined using high-performance liquid chromatography (HPLC) (Thermo Fisher Scientific) equipped with an Aminex HPX-87H column. The mobile phase was 5 mM of H_2SO_4 at a flow rate of 0.5 ml/min. Column temperature was held at 40°C and samples were run for 50 min. Compounds were quantified using Shodex RI-101 and UV/vis (210 nm) detectors. Calibration curves were prepared using standards purchased from Sigma-Aldirch. The concentration of pyoverdine was determined by HPLC (Shimadzu) with a C18 column (4.6 mm × 250 mm) and a UV/vis detector set at 403 nm. The mobile phase contained water, 100 mM formic acid, and acetonitrile (10:10:80 v/v/v) with a flow rate of 1 ml/min at 30 °C. The concentration of pyoverdine was estimated using a standard curve prepared with pyoverdines from *Pseudomonas fluorescens* (>90%) (Sigma-Aldrich). Pyoverdine was additionally detected by fluorescence measured using a Synergy MX reader with excitation and emission wavelengths of 405 nm and 460 nm [295].

Rate calculations

Consumption and production rates at steady state were calculated using differences in concentrations in the liquid and the reactor at steady state:

$$q_s = D \cdot \frac{c_{s,in} - c_{s,out}}{CDW}, \tag{6.1}$$

where q_s is the glucose uptake rate in mmol/g_{DW}/h, $c_{s,in}$ is glucose concentration in the media entering the reactor (mmol/l), $c_{s,out}$ is glucose concentration in the media leaving the reactor at steady state (mmol/l), CDW is cell concentration in the reactor (g_{DW}/l) and D is the dilution rate (h^{-1}).

Consumption and production of O_2 and CO_2 respectively, were calculated using:

$$q_i = \frac{1}{CDW * V_L} \frac{F_{g,out} * \%c_{i,out} - F_{g,in} * \%c_{i,in}}{100} \tag{6.2}$$

where $c_{i,in}$, and $c_{i,out}$ are the O_2 or CO_2 concentrations in the gas (%), $F_{g,out}$, and $F_{g,in}$ are the gas flow rates (mmol/h), and V_L is the liquid volume.

C-recoveries were calculated by dividing the carbon entering the reactor as glucose in the media by the carbon leaving the reactor as CO_2 or other products and assimilated by the cells. A percentage of 47.35% of carbon in biomass was used [296].

RNA isolation and sequencing

For every steady state, samples for RNA isolation were harvested. The sampled volume was calculated to obtain a total cell concentration equivalent to an OD600 of 15 and transferred to 15 ml Falcon tubes containing 1/5 volume of stop solution (5% phenol in 95% ethanol, 4°C). Samples were vortexed for 15 seconds, incubated for 5 min at 4 °C, and centrifuged (3500 g, 10 min, 4°C). Supernatants were discarded and pellets were frozen in liquid N_2 and stored at -80°C until further analysis [297]. RNA isolation was performed according to the manufacturer's instruction of Maxwell® 16 LEV simplyRNA cells kit (Promega). The purified RNA was measured in a NanoDrop spectrophotometer and stored at -80°C. Samples were sequenced by NovoGEne using Total RNA. RNA-seq results are available at the European Nucleotide Archive (ENA) with study accession number PRJEB72150. Sample accession numbers for reactor 1: air, ERR12535875; 5% oxygen, ERR12535874; 2% oxygen, ERR12535876; air after limitation, ERR12535881. Sample accession numbers for reactor 2: air, ERR12535871; 5% oxygen, ERR12535868; 2% oxygen, ERR12535869; air after limitation, ERR12535870.

Differential expression analysis

RNA-seq reads were pre-processed and cleaned using fastp v0.23.1 with default settings [215]. Ribosomal RNA removal was performed using bbduk from the BBMAp suite v38.79 using ribokmers.fa as reference and k=31 [298].

Read counts were obtained with Kallisto v0.46.0 quant with a bootstrap value of 100 [299] using the publicly available coding sequences of *P. putida* retrieved from NCBI with accession number GCA_000007565.2_ASM756v2 [300]. DESeq2 was used to normalize the read counts and to compute differential expression analysis [301]. Statistical significance of gene expression differences was evaluated using a False Discovery Rate (FDR) < 0.05 and $|\log_2(\text{foldchange})| \geq 0.58$ (corresponding to $\log_2(1.5)$) as thresholds. Variance stabilizing transformation was performed to obtain a matrix of expression data based on the normalized count data. The matrix of expression data was used for hierarchical clustering and principal component analysis (PCA). Additionally, maSigPro was used (using counts=True) to identify genes with significant expression changes along the considered time series [302]. Genes were clustered using hierarchical clustering with correlation as distance metric using the Scipy Python library (v1.6.2).

Preparation of proteomics samples and mass spectrometry analysis

At every steady state, the sample volume for proteomic analysis was calculated to obtain a cell concentration equivalent to an OD600 of 20, transferred to a 15 ml Falcon tube, and centrifuged for 5 min at 4500 g. Pellets were washed with 1 ml PBS at 4°C and centrifuged for 5 min at 4500 g. Pellets were frozen in liquid N_2 and stored at -80°C. Protein extraction was performed via MPLEx [303] as described in Gao et al. [304]. The concentration of extracted proteins was measured by microplate BCA assay (ThermoFisher Scientific 23225). 100 µg proteins per condition were transferred into a new tube and brought to a final volume of 100 µl with 100 mM ammonium carbonate. 200 mM TCEP was added to a final concentration of 10 mM and samples were incubated for 1 h (37°C, 850 rpm). 550 mM iodoacetamide stock solution was added to the sample to reach a final concentration of 50-55 mM and incubated for 30 minutes protected from light at room temperature. The concentration of urea was reduced to less than 2 M with 100 mM ammonium carbonate by 10-fold dilution before trypsin digestion.

1 µg/µl solution trypsin (Pierce 90057) in 0.1% formic acid was added in a ratio of 1:50 w/w (trypsin: protein) and samples were incubated overnight at room temperature. Peptide desalting and cleaning were performed via StageTips ®C18 (EMPORE™ - 3M, 66883-U) as described by Rappsilber et al. [305], and peptides were eluted from the column in 20 µl. TMT labeling was performed based on Zecha et al. [306]. A 1:1 ratio (TMT: peptides) using a concentration of 11.8 mM TMT reagent and 4 g/l peptides was employed. The labeling reaction was incubated for 1 h (25°C, 400 rpm). The reaction was quenched with 1M Tris pH 8 to a final concentration of 50 mM for 15 min (25°C, 400 rpm). Equal amounts (20 µl) of each sample were combined in a new microcentrifuge tube and speedvac to dry the labeled peptide sample. Peptides were resuspended in 0.1% formic acid and cleaned up using StageTips [305]. Labeled peptides were stored on a C18 membrane at -80°C until further analysis.

Chromatography separations and mass spectrometry (MS) analyses were performed on an Easy-nLC™ 1000 coupled to an Orbitrap Exploris™ 480 (Thermo Fisher Scientific). Chromatography separation was performed as described in Feng et al. [307]. Mass spectrometry data were acquired in data-dependent (Cycle Time) mode excluding +1 and peptides with unassigned charges, and including charges up to 5+. Peptide full spectra were recorded from 380 to 1400 m/z on the Orbitrap mass analyzer set at 60 k resolution in profile mode, using an AGC target of 5E4, with Custom maximum injection time, and an exclusion time of 15 s. MS/MS spectra were obtained using HCD fragmentation with a fixed normalized collision energy at 36% and an m/z isolation window of 0.7, resolution was set to 45 k, AGC target was set to Custom and 100% normalized. Data searches were run against the *P. putida* KT2440 UniProt database [308] (comprising both reviewed and unreviewed entries downloaded in November 2022) using standard settings on MaxQuant software (v2.0.3.0) [309]. Raw data is available in Zenodo. Data were filtered in R to remove search results of contaminants, reverse, and only identified by site protein groups. Only protein groups with more than 2 peptides and more than 1 unique peptide per protein group were considered.

Differential abundance analysis

Data were normalized for different sample loading in R. The Perseus software (v2.0.11) [310] was used to perform differential abundance analysis using a permutation-based FDR corrected t-test (500 permutations, FDR=0.05). Analysis of variance (ANOVA) was used to identify significantly abundant proteins in all conditions. Bioconductor package maSigPro (counts=False) was used to identify proteins with significant time profiles [302]. Clustering analysis and heatmaps were performed in Python 3.8.8 using Scipy (v1.6.2) and Seaborn (v0.11.1) libraries.

GO-term enrichment analysis

Coding sequences were functionally annotated with GO terms using eggNOG mapper v2.1.6 (settings: -m diamond –evalue 0.001 –score 60 –pident 40 –query_cover 20 –subject_cover 20 –itype CDS –translate –tax_scope auto –target_orthologs all –go_evidence non-electronic –pfam_realign none –report_orthologs) [311]. GO terms were used to generate annotation files for BinGO. The BinGO Cytoscape app (v3.9.1) was used for GO enrichment analysis using the complete functionally annotated genome of *P. putida* KT2440 or the subset of detected proteins as reference for RNA-seq and proteomics data, respectively [312]. The Benjamini–Hochberg FDR multiple test correction and a significant level of 0.05 were applied.

Results

Physiological response of *P. putida* to oxygen and glucose limitations

The adaptive response of *P. putida* to varying oxygen and glucose levels was studied using duplicate chemostat bioreactors (Figure 6.1). After a batch growth phase, reactors were operated as chemostats with a constant dilution rate, and consequently a growth rate, of 0.1 h^{-1}. During the first stage of the chemostat cultivation, cells were glucose-limited as the dissolved oxygen (DO) was maintained above 30%, and no glucose was detected in the effluent. Subsequently, in the next two phases of the cultivation different degrees of oxygen limitation were implemented (DO = 0%) by reducing the oxygen levels in the gas phase of the reactor to 5% and 2%. During these phases, glucose remained in excess and was not completely depleted. In the final stage of the cultivation, air was reintroduced as the gas phase, reinstating glucose limitation.

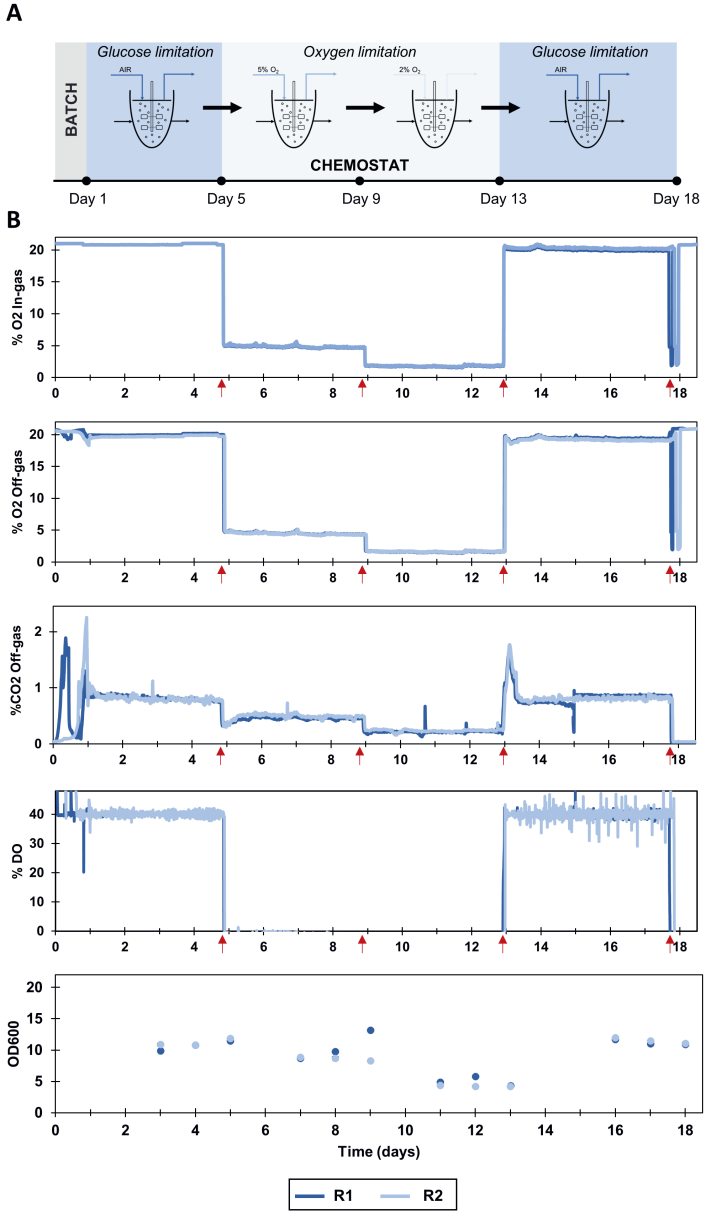


Figure 6.1: Bioreactor setup and measured parameters. **A.** Schematic representation of the experimental setup used for batch and chemostat cultivation. Steady-state samples for physiological characterization, RNA-seq, and proteomic analysis were taken on days 5, 9, 13, and 18 (red arrows). **B.** Online measurements of the duplicate reactors R1 and R2 including oxygen concentration in the in-gas and off-gas, CO_2 concentration in the off-gas, and dissolved oxygen (DO). Duplicate OD600 measurements of the reactor effluent were performed two days before the steady state sampling (standard deviation bars are plotted but are not visible).

During oxygen-limited conditions, cell concentration at steady state decreased while the yield of biomass on glucose increased by 22% and 59% for 5% and 2% oxygen levels, respectively (Figure 6.2, Table 6.1). In the initial glucose-limited phase, no by-products, such as α-ketogluconate, that could explain the reduced biomass yields were identified by HPLC. However, a noticeable yellow color in the supernatant prompted speculation of an overproduction of pyoverdine. This hypothesis was supported by fluorescence measurements (Sup. Figure 6.1) [295]. Despite the absence of a specific *P. putida* pyoverdine standard, the production of this metabolite was further confirmed by HPLC using pyoverdines from *P. fluorescens* as a comparative standard (Figure 6.2, Sup. Figure 6.2). This analysis yielded quantitative estimates that, while approximate, were crucial in closing the carbon mass balances during glucose-limited growth and suggested pyoverdine production as responsible for the observed decrease in biomass yield (Table 6.1). Pyoverdine plays a pivotal role in iron scavenging for *P. putida*, facilitating iron uptake by binding to iron ions with high affinity, and has not been previously reported to accumulate under glucose-limited conditions [313]. According to carbon mass balances, pyoverdine production was not substituted by other by-products during oxygen-limited growth, and all the glucose was directed to biomass synthesis (Table 6.1).

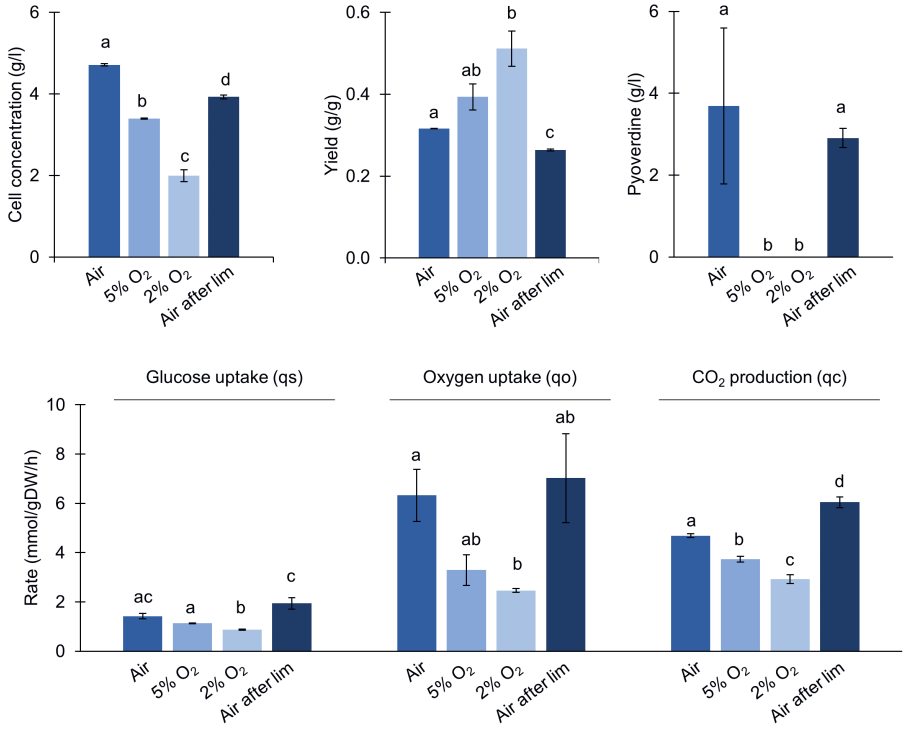


Figure 6.2: Physiological response of *P. putida* under varying oxygen and glucose limiting conditions in a continuous culture. Different letters indicate significant differences between conditions according to two-tailed homoscedastic t-tests.

The calculated glucose uptake obtained during glucose-limited growth (Table 6.1) is in agreement with previous measurements [284]. Upon oxygen limitation, the specific glucose uptake rate was reduced by 21% and 36% when 5% or 2% of oxygen was used. The CO_2 production rate was similarly reduced by 21% and 38%, suggesting that active pathways for glucose metabolism reduced their fluxes but no new pathways were activated in response to oxygen limitation, even when specific oxygen uptake rates decreased by 63% and 60%, respectively (Figure 6.2). When air was reintroduced following the oxygen-limited phase, cells exhibited a significant increase in both glucose uptake and CO_2 production, exceeding the levels observed during the initial glucose-limited phase. Despite this metabolic reactivation, the cell concentration only partially recovered, reaching 83% of its original level. Similarly, the biomass yield did not fully return to its initial efficiency, restoring to only 81% of the yield observed in the initial phase (Figure 6.2).

These results indicated the ability of *P. putida* to endure long-term (up to 8 days) oxygen-limited growth. Upon excess oxygen supply, the cells' physiology exhibited substantial recovery, reaching levels comparable to those observed during the initial glucose-limited phase.

Table 6.1: Physiological parameters of *P. putida* under varying oxygen and glucose limiting conditions in a continuous culture. qs, specific glucose uptake rate; qo, specific oxygen uptake rate; qc, specific CO_2 production rate; Yxs, yield of biomass on glucose. See Sup. Table 1 for additional information.

	Air	**5% O_2**	**2% O_2**	**Air**
qs (mmol/g_{DW}/h)	1.4 ± 0.1	1.1 ± 0.0	0.9 ± 0.0	1.9 ± 0.2
qo (mmol/g_{DW}/h)	6.3 ± 1.1	2.3 ± 0.6	2.5 ± 0.1	7.0 ± 1.8
qc (mmol/$_{DW}$/h)	4.7 ± 0.1	3.7 ± 0.1	2.9 ± 0.2	6.0 ± 0.2
Cells (g_{DW}/l)	4.7 ± 0.0	3.4 ± 0.0	2.0 ± 0.2	3.9 ± 0.1
Yxs (g/g)	0.32 ± 0.00	0.39 ± 0.03	0.51 ± 0.04	0.26 ± 0.00
Pyoverdine (g/l)	3.7 ± 1.9	nd	nd	2.9 ± 0.3
% $C_{recovery}$	121 ± 21	100 ± 4	103 ± 1	100 ± 6

Transcriptional response of *P. putida* to oxygen and glucose limitations

After examining the physiological adaptations of *P. putida* to varying oxygen and glucose levels, we next investigated the transcriptional changes under these conditions to further understand the underlying molecular mechanisms. Samples for RNA-seq were taken at steady states corresponding to cells grown under initial glucose limitation, varying degrees of oxygen limitation (5% or 2% oxygen in the gas phase), and the subsequent restoration of glucose limitation (Figure 6.1A). A total of 4600 transcripts were detected with more than 100 counts in at least one condition, corresponding to 82.7% of *P. putida*'s genes [300].

Principal Component Analysis (PCA) with all the detected counts confirmed the separation of samples based on the degree of oxygen limitation, indicating a transcriptional response to these growth conditions (Sup. Figure 6.3). Differentially expressed (DE) genes were found using

glucose-limited cells at day 5 as reference condition (Sup. Table 2). The most DE genes were found in the comparison with the 2% oxygen condition (896 up-regulations and 877 down-regulations), followed by the second glucose-limited condition (756 up-regulations and 877 down-regulations) and the 5% oxygen condition (611 up-regulations and 376 down-regulations) (Figure 6.3). The unexpectedly high number of DE genes between the two glucose-limited samples (that were taken 13 days apart), as well as the overlap between up- and down-regulated genes between the second glucose-limited samples and cells grown with 2% oxygen, suggested the aging of the culture as possible confounding effect on the pair-wise identification of DE genes.

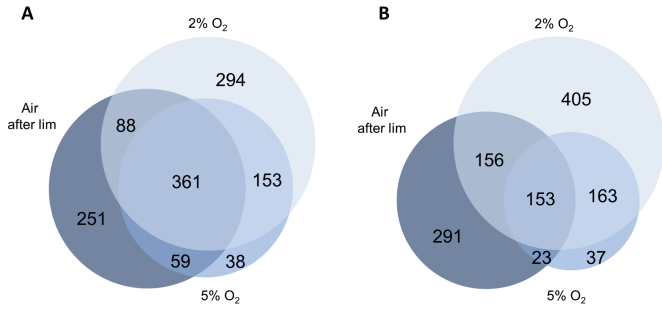


Figure 6.3: Differential expression analysis (see Sup. Table 2 for details). **A.** Euler diagram showing up-regulated genes comparing the indicated conditions to glucose-limited cells at day 5. **B.** Euler diagram showing down-regulated genes comparing the indicated conditions to glucose-limited cells at day 5.

To avoid confounding the effect of oxygen limitation and time on the identification of DE genes, the maSigPro workflow, designed for the analysis of time-course RNA-seq data was applied [302]. This package performs single-time course analysis based on a two-step regression strategy to find genes with significant temporal expression changes [302, 314]. This workflow identified 1806 genes with significant changes in expression over time. Using these genes for clustering led to a clear separation of samples based on the imposed limitation (Sup. Figure 6.4A). While pair-wise DE analysis is unable to distinguish changes in RNA levels caused by the aging of the culture or oxygen limitation, the two-step regression strategy followed by clustering allows the distinction between genes up- or down-regulated due to oxygen-limited growth or to the culture's age (Sup. Figure 6.4B). This refined analysis identified 321 genes up-regulated and 602 genes down-regulated specifically during oxygen-limited growth (Figure 6.4A, C, Sup. Table 3).

Gene ontology (GO) analysis of genes up-regulated during oxygen limitation showed enrichment of functions related to the respiratory chain, specifically iron-sulfur cluster binding, complex I (NADH dehydrogenase), and cytochrome complex assembly (Figure 6.4B). Genes down-regulated during oxygen limitation were enriched on catabolic processes including fatty acid β-oxidation and protein maturation (Figure 6.4D). This suggests that, at the transcriptional level, cells respond to oxygen limitation by over-expressing genes related to respiration while only halting a few catabolic processes.

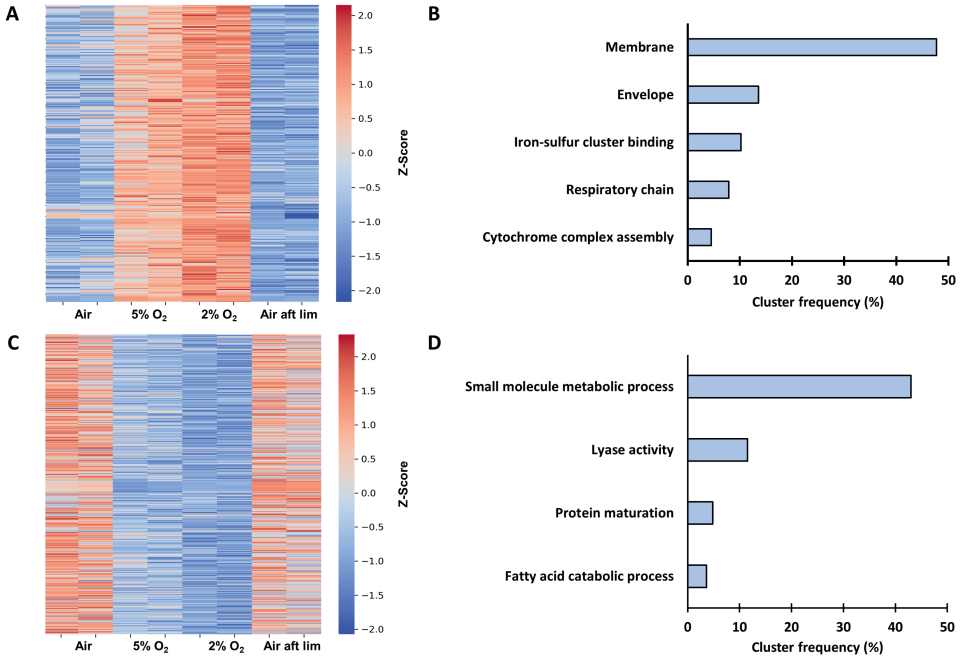


Figure 6.4: Two-step regression analysis of RNA-seq data. **A.** Heat-map of z-scores of 321 genes up-regulated during oxygen-limited compared to glucose-limited growth. **B.** Frequency of selected GO-terms enriched in genes up-regulated during oxygen limitation. **C.** Heat-map of z-scores of 602 genes down-regulated during oxygen-limited compared to glucose-limited growth. **D.** Frequency of selected GO-terms enriched in genes down-regulated during oxygen limitation.

Proteomic-level response of *P. putida* to oxygen and glucose limitations

The extent to which the observed transcriptional changes affected *P. putida*'s physiology was evaluated using proteomics. Samples for proteomic analysis corresponding to the steady state conditions of glucose-limited cells, cells treated with 5% oxygen, 2% oxygen, and glucose-limited cells after oxygen limitation were analyzed. A total of 1036 proteins were quantified via MS-based proteomics analysis in all conditions, representing 22.5% of the detected transcripts. A comparison between changes in protein abundance and their transcripts is shown in Sup. Figure 6.5.

First, detected proteins were compared pairwise among all tested conditions to identify differential abundant (DA) proteins. This is equivalent to the initial DE analysis of the RNA-seq data. DA analysis only found significant differences when comparing initial glucose-limited cells or cells treated with 5% or 2% oxygen with samples from the second glucose-limited period (Sup. Figure 6.6, Sup. Table 4). Although 43 DA proteins were found comparing the initial and final glucose-limited conditions, these proteins were not enriched in any specific GO term. When comparing cells grown in oxygen limitation with the final glucose limitation, an enrichment of GO terms related to aerobic respiration was found, in agreement with the RNA-seq results. GO terms enriched

among the down-regulated genes during oxygen-limited growth were related to gene expression, ribosome assembly, and peptide synthesis.

Considering that no different protein abundances were found when comparing initial glucose-limited cells and oxygen-limited cells, ANOVA was used to find DA proteins among all conditions simultaneously. This resulted in the identification of 9 proteins. When these proteins were used for clustering, samples from the first and second glucose limitation were separated and samples from oxygen-limited growth were clustered together (Sup. Figure 6.7A). Heatmaps were used to distinguish proteins with different abundance due to oxygen limitation or time (Sup. Figure 6.7B). Two of the DA proteins (PP_0842 and PP_4220) showed decreased abundance during oxygen limitation (Table 6.2, Sup. Figure 6.7B). PP_0842 (IscS-I) is a cysteine desulfurase involved in iron-sulfur cluster synthesis, characteristic of ferredoxins involved in the electron transport chain [315]. PP_4220 (PvdJ) is a non-ribosomal peptide synthase involved in pyoverdine synthesis, only observed during glucose-limited growth [316]. The only protein more abundant during oxygen limitation was PP_3839 (CalA), an alcohol dehydrogenase previously known as AdhP (Table 6.2, Sup. Figure 6.7B) [317].

Table 6.2: Proteins with increased or decreased abundance during oxygen-limited growth found using ANOVA and two-step regression analysis.

Gene	Gene name	Oxygen limitation	Function
PP_0842	*iscS-I*	Down	Fe-S cluster synthesis
PP_4220	*pvdJ*	Down	Pyoverdine synthesis
PP_3839	*calA*	Up	Alochol dehydrogenase
PP_0625	*clpB*	Up	Chapeorne
PP_5392	-	Up	β-propeller fold lactonase
PP_0481	*katA*	Up	Catalase
PP_1084	*tsaA*	Up	Peroxidase

In order to identify proteins with significant time-dependent changes in abundance, the two-step regression analysis applied to the RNA-seq data was also applied to the normalized proteomic data using maSigPro [314]. This strategy resulted in the identification of 18 proteins. When these proteins were used for clustering, they were able to separate samples based on their nutrient limitation, with one cluster containing samples from glucose-limited cells and another cluster containing samples corresponding to oxygen-limitation (Sup. Figure 6.7C). Heatmaps were used to distinguish proteins with abundant changes caused by oxygen limitation or time (Sup. Figure 6.7D). Four proteins were more abundant during oxygen-limited growth: PP_0625 (ClpB), PP_5392, PP_0481 (KatA), and PP_1084 (TsaA) (Table 6.2, Sup. Figure 6.7D). These proteins have chaperone, β-propeller fold lactonase, catalase, and thioredoxin peroxidase functions [279, 318, 319]. None of the proteins found using this approach showed a decreased abundance specific to oxygen-limited conditions (Sup. Figure 6.7D).

Discussion

The study of bacterial physiology and metabolism traditionally focuses on conditions of exponential growth, characterized by the abundance of nutrients. However, this ideal scenario often diverges significantly from the heterogeneous environments encountered in industrial bioprocesses, especially in large-scale bioreactors. In these practical settings, microorganisms like *P. putida* are subjected to a dynamic landscape of nutrient availability, often resulting in gradients and limitations that markedly influence their physiological state [283, 291]. Furthermore, in fed-batch-operated reactors, cells exhibit controlled growth, often below the maximum growth rate. These sub-optimal conditions, particularly concerning key nutrients like glucose and oxygen, necessitate a deeper exploration into how microbial cells adapt and respond. Such understanding is crucial not only for optimizing growth and production but also for ensuring stability and consistency in bioprocesses. Our research, therefore, delves into the physiological adaptations of *P. putida* to varying degrees of oxygen limitation, using glucose limitation as a reference. This approach sheds light on the cellular response mechanisms to nutrient stresses and also provides valuable insights for enhancing the efficiency and robustness of microbial processes in industrial applications.

In this study, *P. putida* displayed a maximum biomass yield on glucose (0.40 g/g) under 5% oxygen, similar to yields in rapid growth scenarios [284]. Although the yield of biomass decreases at low growth rates, this decrease was previously attributed to changes in cell size [284]. However, utilizing glucose-limited cells as a reference to investigate *P. putida*'s response to oxygen limitation uncovered pyoverdine production as a factor influencing low yields under slow, glucose-limited growth. Pyoverdine, a siderophore used for iron capture, is known to be secreted in iron-limited media [282, 313, 320]. Additionally, the regulation of pyoverdine synthesis has been linked to glucose metabolism, with gluconate accumulation stimulating pyoverdine synthesis [321]. However, pyoverdine production by *P. putida* under iron-replete conditions has not been previously reported [282, 283, 284]. In contrast to studies comparing exponentially growing cells in media with and without iron excess or repeated glucose shortage, we studied the physiology of *P. putida* under prolonged, glucose-limited slow growth (0.1 h^{-1}). We show how these conditions, equivalent to glucose-limited fed-batch fermentations, can lead to up to a 59% decrease in biomass yield due to pyoverdine production. Upon oxygen limitation, an increase in biomass yield was observed alongside with the elimination of pyoverdine production and a decline in PvdJ protein abundance, which together with PvdD and PvdI function as non-ribosomal peptide synthases of the pyoverdine peptide side chain [316]. The absence of pyoverdine production under oxygen limitation, as reported by Lenhoff, can be attributed to the oxygen requirements for siderophore production [320]. Moreover, low oxygen tensions increase iron bioavailability and the *Enterobacteriaceae* family regulates iron transport genes using oxygen-sensing regulators [322]. While the mechanisms leading to pyoverdine production under glucose-limited growth remain unclear, pyoverdines produced by fluorescent pseudomonads facilitate their colonization of different hosts [313, 323] and could play a similar role in the colonization of environments where carbon is scarce.

We used transcriptomics and proteomics to further understand the responses of *P. putida* to oxygen and glucose limitation. Although, as commonly observed, a clear correlation between changes in protein abundance and changes in RNA counts was not found for most of the detected proteins [324], the same expression and abundance patterns were present for proteins with significant changes (Sup. Figure 6.5). Besides, in agreement with Demling et. al. who studied the response of *P. putida* to oscillations in oxygen tension [291], we only found a few proteins related to adaptations to oxygen-limited growth (Table 6.2). This contrasts with the adaptation of *P. putida* to other stressors and nutrient limitations, where greater changes at the proteome level have been observed [280, 281, 325], indicating that *P. putida* is able to endure long-term oxygen limitation with minor changes in its proteome.

The approach followed in this study, where chemostat reactors were maintained for various steady states under different conditions, aimed to reduce biological variability when comparing conditions with different limitations. However, this setup introduced time as a confounded factor during analysis, and changes related to the aging of the culture or the change in conditions had to be discerned. This was possible by identifying genes with significant temporal expression changes using a single time course analysis based on a two-step regression strategy [302, 314]. Applying this approach to RNA-seq data allowed the simultaneous comparison of all the conditions and reduced the number of significant genes compared to DE analysis using glucose limitation as reference. While DE genes unique to the comparison between glucose limitation and growth with 2% oxygen were not enriched in any GO term, genes found using maSigPro were enriched in GO terms related to respiration. Although this approach reduced the number of significant genes, it increased the number of significant proteins compared to the two-way comparison, probably due to the removal of confounding effects.

At the transcriptome level, *P. putida* responds to oxygen limitation by over-expressing genes related to respiration. A similar response has been observed in *Saccharomyces cerevisiae* cells in the transition from aerobiosis to anaerobiosis and suggests that cells respond to low oxygen tensions aiming to capture the available oxygen and to maintain respiration [326] . Although an increased abundance of respiration-related proteins was not detected, this could be attributed to the difficulty in detecting and quantifying membrane proteins in proteomic studies [327, 328]. Instead, a significantly increased abundance of CalA, ClpB, PP_5392, KatA, and TsaA was observed. While over-expression of *clpB* and *tsaA* has been reported as a response to formaldehyde [317] and poor carbon sources like phenol or pyruvate [319], *calA*, previously known as *adhP*, was over-expressed when *P. putida* cells were exposed to temporary oxygen limitation [291]. Although the increased abundance of KatA during oxygen limitation is counterintuitive as it is used to detoxify oxygen reactive molecules [329], it has been described to prepare *Staphylococcus aureus* for future oxidative stress under microoxic conditions [318]. However, the over-production of these proteins does not suggest a specific *P. putida* response to low oxygen tensions.

Contrarily to facultative anaerobic microorganisms, the onset of oxygen limitation does not result in the production of fermentative products by *P. putida*, preventing the loss of energy and

carbon towards unwanted metabolites. Although *P. putida* possesses genes associated with low-oxygen metabolism, including a lactate dehydrogenase (PP_1649), formaldehyde dehydrogenases (PP_4960, PP_0328, PP_3970), and an acetoin gene cluster (PP_0550-0556), none of these genes are up-regulated during oxygen-limited growth, and their associated proteins are not detected. This finding aligns with the observation from Ankenbauer et al. that excess electrons in such conditions can be redirected towards the production of reduced products [292]. Moreover, no major changes in metabolic genes were observed during oxygen-limited growth, and only genes related to β-oxidation showed a decreased expression that was not observed at the protein level. The only detected proteins with a decreased abundance during oxygen limitation were PvdJ and IscS-I. While the decreased abundance of PvdJ agrees with the lack of pyoverdine production in this condition, IscS-I is involved in the reparation of FeS clusters, which are sensitive to oxygen, and less prone to damage during oxygen limitation [315, 330]. Therefore, if *P. putida* is grown under microoxic conditions, major metabolic changes are not expected and a decreased availability of precursors for product synthesis is unlikely.

Our research demonstrates the resilience of *P. putida* to long-term oxygen limitation, revealing a notable increase in biomass yield without significant metabolic reconfigurations. This adaptation, beneficial for bioprocess efficiency, comes with a trade-off: reduced glucose uptake rates under oxygen limitation. This reduction in glucose uptake implies a lower metabolic rate, which could affect the synthesis of desired products, particularly in processes where high productivity metrics are crucial [160]. Despite this, growing *P. putida* in microoxic conditions reduces energy demands for aeration and cooling in industrial bioreactors and enhances the production of reduced compounds. The discovery of pyoverdine production in glucose-limited, slow-growing cultures further underscores the complexity of microbial metabolism and the importance of thorough mass balance analyses for bioprocess optimization. This finding suggests an adaptive response to nutrient availability and highlights the need to evaluate biotechnological organisms like *P. putida* under diverse conditions for industrial applicability. In summary, our study contributes to a better understanding of *P. putida*'s physiological responses, offering insights for future research aimed at balancing yield and productivity in biotechnological applications, ultimately leading to more efficient and sustainable industrial processes.

Declaration of interest

Vitor A. P. Martins dos Santos has interests in LifeGlimmer GmbH.

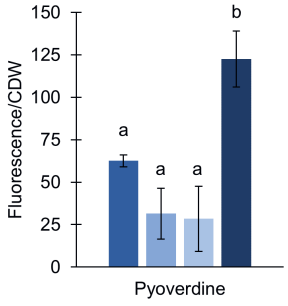
Acknowledgment

This project was funded by NWO (project number GSGT.2019.008). Mass spectrometry data were acquired at the Laboratory of Biochemistry, Wageningen University and Research. We would like to acknowledge Tom Schonewille, Silvia Rodriguez Marcos and Bart Nijsse for their valuable help in establishing the reactor setup,the extraction of proteomic samples and the analysis of RNA-seq data, respectively.

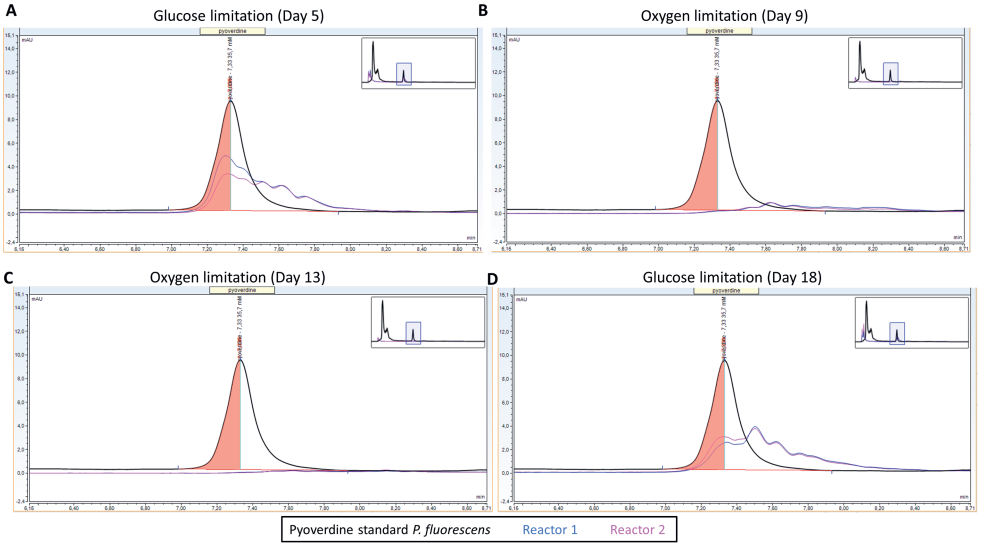
Data availability

RNA-seq data is available at ENA with accession number PRJEB72150. Raw proteomic data and Sup. Tables including physiological data, significant genes found using DE analysis and maSigPro and significant proteins found using ANOVA and maSigPro are available at Zenodo.

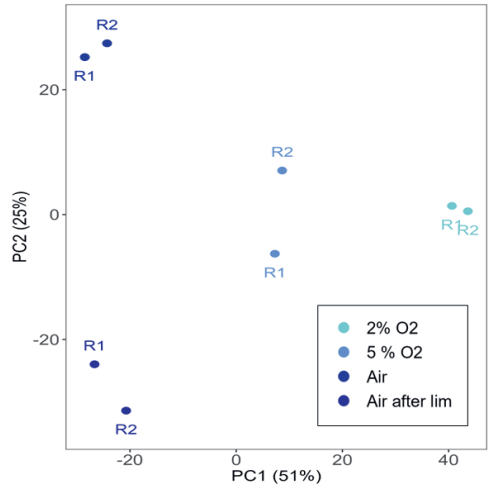
Supplementary Figures



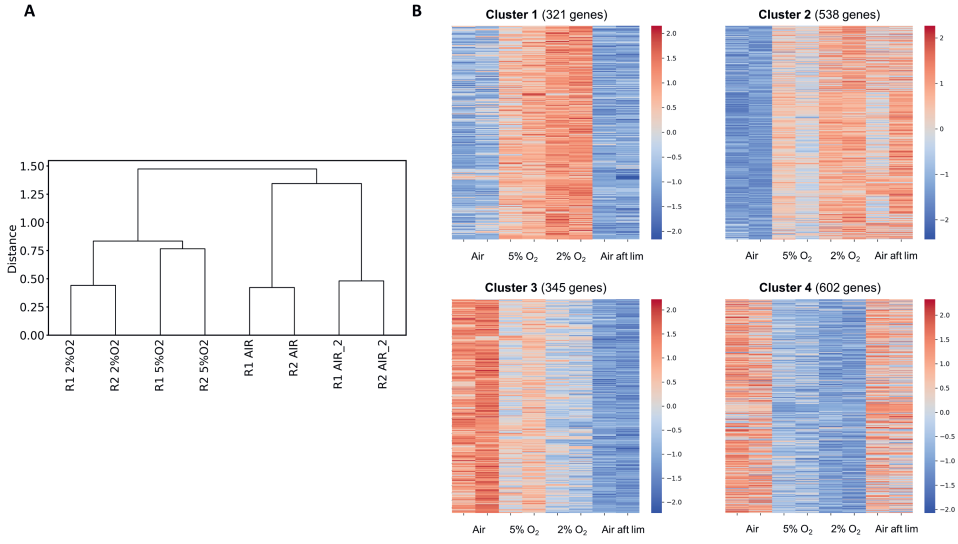
Sup. Figure 6.1: Fluorescence of samples normalized by dry weight. Different letters indicate significant differences between conditions according to two-tailed homoscedastic t-tests.



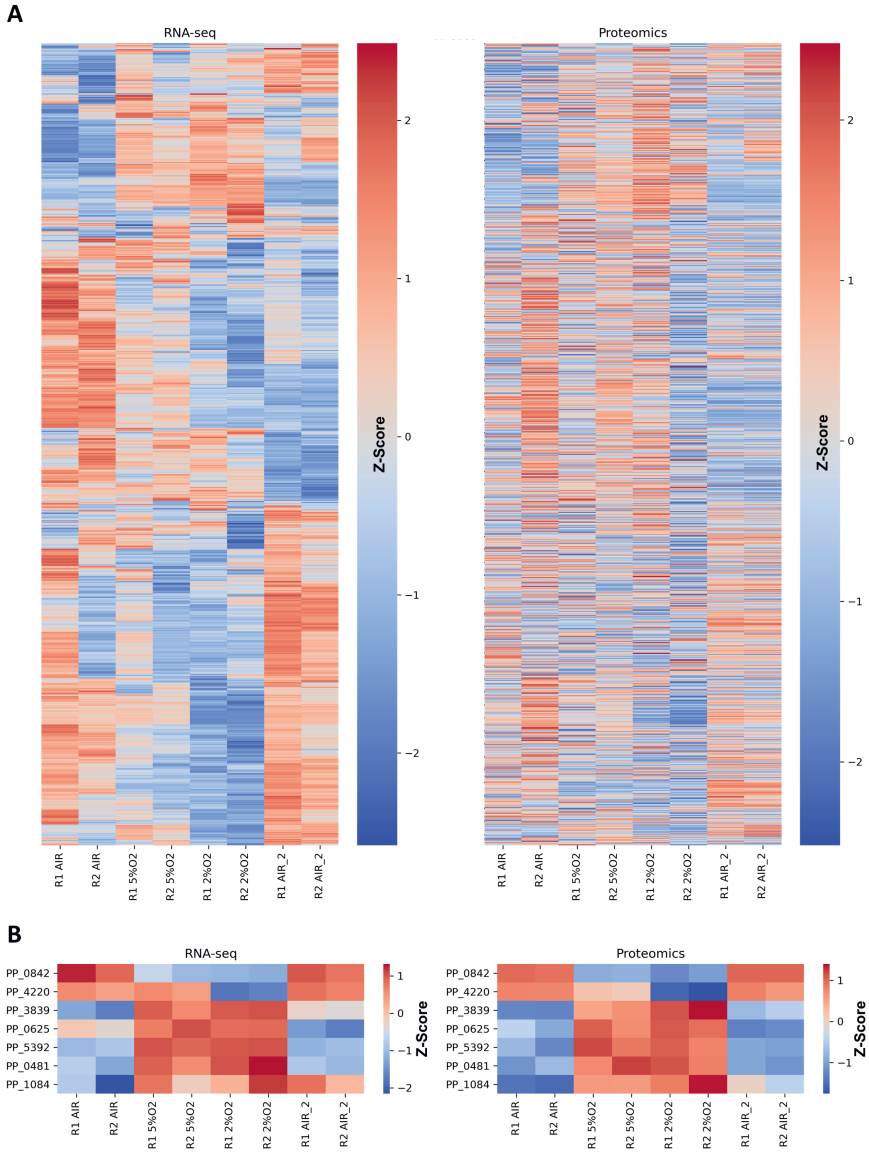
Sup. Figure 6.2: Chromatograms used for pyoverdine identification.



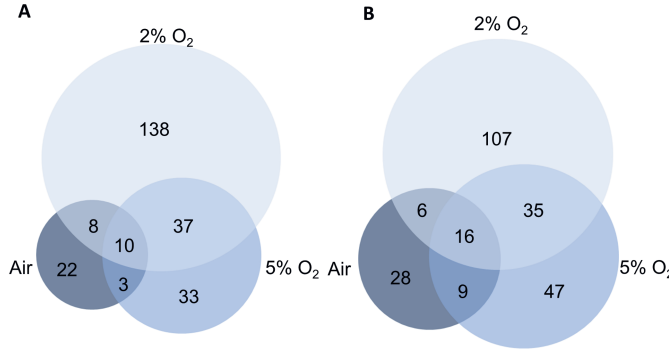
Sup. Figure 6.3: PCA analysis with RNA-seq data. PC1, explaining 51% of the variance, separates samples based on the degree of oxygen limitation, indicating a transcriptomic response to this growth conditions. PC2, which explains 25% of the variance, separated samples from glucose-limited cells before and after oxygen limitation, suggesting a possible transcriptomic-level effect caused by long-term oxygen limitation or the aging of the culture. R1 and R2 refer to reactors 1 and 2.



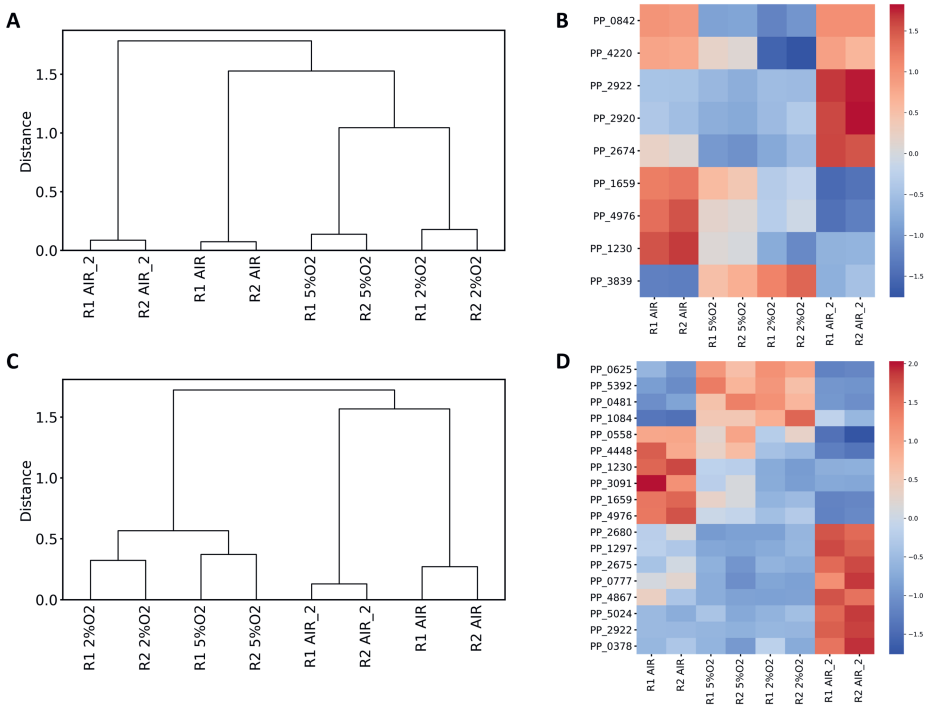
Sup. Figure 6.4: Two-step regression analysis of RNA-seq data. **A.** Dendogram showing the clustering of samples according to the nutrient limitation applied based on the 1806 significant genes found with maSigPro. **B.** Heatmaps of the z-scores of the 1806 significant genes that were divided into 4 clusters according to the changes in expression. Cluster 1 contains genes that were up-regulated during oxygen-limited growth, cluster 2 contains genes up-regulated with time, cluster 3 contains genes down-regulated with time, and cluster 4 contains genes down-regulated during oxygen-limited growth. Genes belonging to each cluster are available in Sup. Table 3.



Sup. Figure 6.5: **A.**Comparison of the abundance of detected proteins with their respecitive transcripts. Heatmaps are clustered based on RNA-seq data. **B.** Zoom in only including differentially abundant proteins.



Sup. Figure 6.6: Differentially abundant proteins based on pair-wise comparison using the second glucose-limited sample (day 18) as reference (See Sup. Table 4 for details). **A.** Proteins more abundant compared to the reference condition. **B.** Proteins less abundant compared to the reference condition.



Sup. Figure 6.7: Significant proteins found using ANOVA and two-step regression. **A.** Dendogram of samples based on significant proteins found using ANOVA. **B.** Heat-maps of the z-score of the 9 significant proteins found using ANOVA. **C.** Dendogram of samples based on significant proteins found using two-step regression analysis. **D.** Heat-map of the z-score of the 18 significant proteins found using two-step regression analysis.

CHAPTER 7

In silico analysis of Design of Experiments for metabolic pathway optimization

Sara Moreno Paz, Joep Schmitz, María Suárez Diez

This chapter is accepted for publication in *Computational and Structural Biotechnology Journal*

Abstract

Finding the optimal expression of production pathway genes is crucial for the development of efficient production strains. Unlike sequential experimentation, combinatorial optimization captures the relationships between pathway genes and production, albeit at the cost of conducting multiple experiments. Fractional factorial designs followed by linear modeling and statistical analysis reduce the experimental workload while maximizing the information gained during experimentation. However, guidelines for selecting appropriate factorial designs for pathway optimization are missing. In this study, we leverage a kinetic model of a seven-genes pathway to simulate the performance of a full factorial strain library. We compare this approach to resolution V, IV, III, and Plackett Burman (PB) designs. Additionally, we evaluate the performance of these designs as training sets for a random forest algorithm aimed at identifying best-producing strains. Evaluating the robustness of these designs to noise and missing data, traits inherent to biological datasets, we find that while resolution V designs capture most information present in full factorial data, they necessitate the construction of a large number of strains. On the other hand, resolution III and PB designs fall short in identifying optimal strains and miss relevant information. Besides, given the small number of experiments required for the optimization of a pathway with seven genes, linear models outperform random forest. Consequently, we propose the use of resolution IV designs followed by linear modeling in Design-Build-Test-Learn (DBTL) cycles targeting the screening of multiple factors. These designs enable the identification of optimal strains and provide valuable guidance for subsequent optimization cycles.

Introduction

A common challenge when introducing a heterologous pathway in a microorganism is to find the optimal expression level of each of the introduced genes [32, 87, 331]. This question can be answered using sequential or combinatorial experimentation, depending on whether the expression of the genes is optimized individually (one factor at a time) or simultaneously. When combinatorial optimization is used, the likelihood of finding the optimal expression levels increases [73, 332]. For example, if the abundance of protein A is limiting the pathway, the expression of other pathway genes will not affect production as long as the expression of A is low. However, when the expression of gene A increases, changes in the expression of other pathways genes will likely affect production. Combinatorial pathway optimization captures these interactions between the pathway genes and can better guide the pathway optimization process.

Combinatorial optimization requires the construction of numerous strains. When optimizing a pathway with three genes and two expression levels, constructing eight (2^3) strains is needed to test all the combinations of genes and levels (full factorial design). The number of strains to construct increases exponentially with the number of genes to optimize: if the number of genes increases to seven, the number of strains increases to 128 (2^7). Moreover, the number of strains to build increases even faster when more than two expression levels are tested (*e.g.* 27 (3^3) strains for three genes with three expression levels and 2187 (3^7) for seven genes with three expression levels). Even with efficient and automated strain construction and characterization pipelines, reducing the number of strains to build and test while maintaining the ability to discern the relative importance of the pathway genes and the presence of interactions is desired [73, 332].

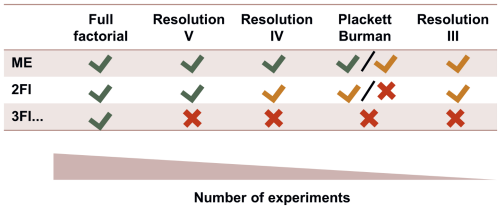
Statistical design of experiments (DoE) is a technique to minimize the experimental effort while the information gained over the studied system is maximized [72]. This method can be easily incorporated in the Design-Build-Test-Learn (DBTL) cycles commonly used in industrial biotechnology [32, 87]. In the first round, factors (*e.g.* genes whose expression will be optimized) and levels (*e.g.* the expression levels that will be tested) are defined. Considering that the number of experiments to perform (*i.e.* strains to build and test) increases faster with the number of levels than the number of factors, DoE often starts studying two levels per factor. In this way, more factors can be screened, important factors are identified, and the fine-tuning of factor levels is targeted only to the relevant factors in subsequent DBTL cycles. The information gained during experimentation is stored in a polynomial model. The model coefficients are fit such as the response (*e.g.* production of the target molecule) is a function of each of the factors and their interactions. In this model, the main effects are the coefficients that explain how the response is affected by changing each individual factor (gene). Similarly, two-factor interactions are coefficients that explain how the response changes simultaneously considering the levels of two factors (genes). Then, an analysis of variance (ANOVA) is used to quantify whether each model parameter significantly influences the response [72, 73]. The extent to which a significant factor influences the response is determined by the absolute value of its main effect, and the sign of the coefficient indicates whether the factor has a positive or negative impact on the response. When two or multiple factor interactions are

significant, the effect of a factor is also influenced by the interaction coefficients.

Factorial designs are used to efficiently sample the design space determined by the factors and their levels and are useful for screening in initial DBTL cycles. Different factorial designs exist depending on the number of experiments to perform and the aliasing structure (*i.e.* which model coefficients are indistinguishable from each other). A design with a higher resolution requires the execution of more experiments and results in the confounding of only high-order interactions [72]. For instance, resolution V designs allow the clear identification of main effects and two-factor interactions while confounding three-order interactions among each other. This means that, although some three-factor interaction coefficients can be estimated, they cannot be assigned to a specific combination of factors. Similarly, resolution IV designs confound two-factor interactions among each other. Therefore, they can be used to assess whether these interactions are important but they cannot identify the specific interactions that influence the response. Resolution III designs confound main effects with two-factor interactions so, although models including main effects can be created, the estimated coefficients represent the mixed effect of the single factor and the confounded interactions. If interactions are not significant, low-resolution designs efficiently reduce the number of experiments, but they may result in incorrect determination of main effects when interactions affect the response. Plackett Buramn (PB) designs are a special type of resolution III designs in which two-factor interactions are partially confounded with main effects allowing the estimation of some interactions. A summary of DoE designs is presented in Figure 7.1 [72]. Although factorial designs have been used for the optimization of expression of pathways genes [86, 87], clear recommendations of the type of designs to use for this application are missing [51].

In silico studies represent biological systems using mathematical models which allows the simulation of multiple constructs and the evaluation of computational design tools [333, 334, 335]. These studies enable the characterization of the robustness of different design approaches to realistic biological scenarios where noise is present and problems during strain construction can lead to the inability to build some of the desired strains. The best strategies found by an *in silico* evaluation can then be applied in *in vivo* studies, in which the experimental throughput is considerably lower.

Here, we use a mathematical kinetic model of the curcumin pathway (Figure 7.2A) to simulate *in silico* a full factorial library consisting of all the combinations of seven enzymes (factors) at two different concentrations (levels) [336]. This pathway is characterized by the presence of promiscuous enzymes that catalyze multiple reactions and the possibility to produce three different metabolites. Therefore the effect of modifying the abundance of an enzyme on production is highly dependent on metabolite concentrations affected, in turn, by the concentration of other pathway enzymes. The use of a kinetic pathway model enabled the identification of the best concentration levels of each enzyme, as well as the estimation of the real coefficients of the polynomial model. Considering this information, we tested the capacity of different factorial designs to find the best strains in the library space, as well as to determine the coefficients of the model which could later be used to guide the expansion of the design space. We also provide some recommendations on how to approach the subsequent DBTL cycles.



# Factors	Full factorial	Resolution V	Resolution IV	Plackett Burman	Resolution III
2	4	4	4	-	4
3	8	8	8	8	4
4	16	16	8	8	8
5	32	16	16	12	8
6	64	32	16	12	8
7	128	64	16	12	8
8	256	64	16	12	16
9	512	128	32	12-28	16
10	1024	128	32	12-28	16

Figure 7.1: Ability of different fractional factorial designs to estimate main effects (ME), two-factor interactions (2FI), and higher-order interactions (3FI). A green tick indicates the estimation of a coefficient clear of confounding, an orange tick indicates a confounded estimation and a red cross indicates the inability to estimate this type of coefficient. In Plackett Burman designs confounded coefficients are partially correlated with each other, which allows the estimation of some of the interactions.

Material and Methods

Pathway simulation and noise

A kinetic model of the curcuminoid pathway was obtained from Martín-Pascual et al. [336]. This model uses Michaelis-Menten kinetic expression rate laws for all reactions except C3H, for which a mass-action rate law is used. Enzymes catalyzing multiple reactions (FCS, DCS, and CURS) contain additional substrate competition terms in their rate laws to account for their promiscuity (Figure 7.2A). The model was simulated using the AMICI library [136] and the CVODES ODE solver [135]. Each of the seven enzymes in the pathway was considered a factor with the default enzyme concentration as *low* level and five times the default concentration as *high* level. The parameter corresponding to enzyme concentration for each reaction in the pathway was altered to simulate the 128 (2^7) strains constituting the full factorial library. *In silico* triplicates for each strain were obtained adding 5% or 20% of Gaussian noise. The simulated full factorial data is available in Gitlab.

Simulation of DoE designs

Resolution V, IV, and III designs were generated using the FrF2 function from the FrF2 R package given the number of factors and the desired resolution [337]. Placket-Burman (PB) designs were generated with the pb function from the same package indicating the desired number of factors

and experiments. For each design, columns were permuted to account for the effect of randomly assigning factors (enzymes) to the design columns. From the full factorial design data, experiments were selected according to the design and used to train a linear model by ordinary least squares regression using the R lm function. The linear model had the form:

$$y = \beta_0 + \sum_{i=1}^{i=n} ME_i \cdot F_i + \sum_{i=1}^{i=n} \sum_{j=1}^{j=n} 2FI_{i:j} \cdot F_i \cdot F_j, \tag{7.1}$$

where y represents the curcumin concentration obtained by the kinetic pathway model; β_0 represents the y-intercept, ME_i refers to the main effect of factor i (F_i) and $2FI_{i:j}$ refers to the two-factor interaction between factor i and j. The total number of factors is indicated by n.

For resolution V and IV designs, additional linear models were trained assuming the inability to construct some of the proposed strains by randomly removing rows of the design matrix. The effect of excluding 1, 2, 5, and 10 rows or 1, 2, 3, and 5 rows for the resolution V and IV designs respectively was evaluated in 100 random permutations of the design columns.

To compare DoE designs with random sampling strategies, experiments from the full factorial design were randomly sampled using the sample function in R. The number of samples was equal to the number of strains selected by each of the designs and the sampling process was repeated as many times as the performed permutations.

For each permutation of the design or random sample, the R summary function was used to obtain the ANOVA table which provides the estimated coefficients of the linear model (MEs and 2FIs) and their associated p-values. These p-values and coefficients were compared to those obtained by training the linear model with the full factorial design ("ground truth").

Prediction of optimal strains within the design space

Linear models trained with data derived from experiments selected by permutations of DoE designs or random sampling were used to predict strains with the highest curcumin production. Strains were predicted as optimal producers according to the linear models when the levels of the enzymes with a significant main effect agreed with the sign of the estimated coefficients in the linear model. For enzymes with insignificant main effects, strains containing any of the concentration levels were considered as optimal candidates. The frequency in which each strain was selected as optimal in each permutation of the design or set of random samples was computed and compared to the actual production according to the kinetic pathway model.

DoE and machine learning

The suitability of experiments designed using DoE to train machine learning (ML) models was assessed with random forest as an example using the scikit-learn Python library. Models were trained using 10-fold cross-validation and model performance was assessed based on the coefficient of determination (R^2). Trained models were used to predict the production of the full factorial design space and the frequency of each strain as part of the two best predicted strains was computed.

Results

Simulation of the full factorial library and factorial designs

The curcumin pathway contains seven enzymes from which FCS, DCS, and CURS are promiscuous and able to catalyze multiple reactions (Figure 7.2A). Moreover, in this pathway demethoxycurcumin, bisdemethoxycurcumin, and curcumin can be produced. Therefore, optimizing curcumin production requires fine-tuning the concentration of the pathway enzymes. A full factorial library for this pathway considering two concentration levels per enzyme requires the simulation of 128 strains. This library contains all possible combinations between factor levels, resulting in curcumin production ranging from 10^{-4} to 0.2 mM, and represents the "ground truth" for the system (Figure 7.2B). In factorial designs, the effect of a factor is not only estimated considering replicate experiments but also all the experiments where the given factor is constant regardless of other factor levels. For each enzyme, Figure 7.2A shows the distribution of curcumin production by strains containing low (-1) or high (1) enzyme concentrations. Changing the concentration of C3H, COMT, and FCS has the highest impact on curcumin production, followed by changes in CCOAOMT and CURS concentrations. Notably, as expected for biological systems, high expression of all pathways genes does not necessarily result in optimal production.

The benefit of combinatorial experimentation compared to sequential experimentation is exemplified in Figure 7.2C. When the concentration of the COMT enzyme changes given a low concentration of C3H, changing COMT expression has a limited impact on production. However, this impact increases when the concentration of C3H is high. The relationship between COMT and C3H, also true for other enzymes (Sup. Figure 7.1), can only be captured through combinatorial optimization and is missed when factors are optimized sequentially.

After the simulation of the full factorial library, DoE fractional factorial designs were simulated selecting experiments according to four different designs: resolution V, resolution IV, resolution III, and PB, or random sampling. These designs differ in the number of strains to build and test (as indicated in Figure 7.1 for experiments with 7 factors). The resolution V design requires the construction of 64 strains and ensures that main effects and two-factor interactions are free of confounding. The resolution IV design requires the construction of 16 strains but two-factor interactions are confounded among each other. Finally, PB and resolution III designs require the construction of 12 and 8 strains respectively but confound main effects with two-factor interactions. In resolution III designs main effects and two-factor interactions are completely confounded and, in PB designs, the correlation between these coefficients is partial. While in random designs any of the strains can be constructed, only a fraction of the strains are selected in DoE fractional designs to ensure orthogonality in the desired columns that allows a clear estimation of the linear model coefficients (Sup. Figure 7.2).

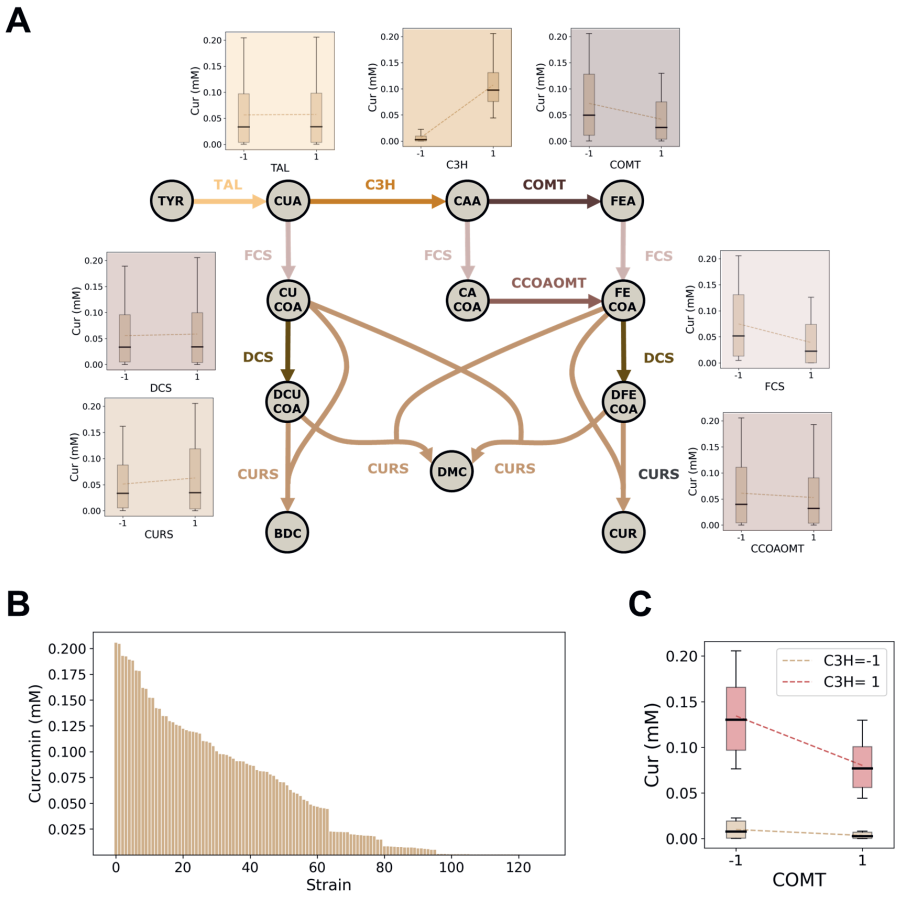


Figure 7.2: **A.** Curcuminoid pathway. Boxplots show the curcumin (Cur) production data distribution of strains containing low (-1) or high (1) enzyme concentration obtained with the kinetic model [336]. **B.** Kinetic model simulations of curcumin production of the 128 strains forming the full factorial design space. **C.** Production data distribution of strains containing low (-1) or high (1) concentration of COMT given low (-1) or high (1) concentration of C3H. Metabolite abbreviations: TYR, tyrosine; CUA, p-coumaric acid; CAA, caffeic acid; FEA, ferulic acid; CUCOA, coumaroyl-CoA; CACOA, caffeoyl-CoA; FECOA, feruloyl-CoA; DCUCOA, diketide coumaroyl-CoA; DFECOA, diketide feruloyl-CoA; BDC, bisdemethoxycurcumin; DMC, demethoxycurcumin; CUR, curcumin. Enzyme abbreviations: TAL, tyrosine ammonia lyase; C3H, coumarate-3-hydroxylase; COMT, caffeic acid O-methyl transferase; FCS, feruloyl/coumaroyl-CoA synthase; CCOAOMT, caffeoyl-CoA O-methyl transferase; DCS, diketide-CoA synthase; CURS, curcumin synthase.

Pathway optimization: predictions of optimal strains

Given a set of enzymes (factors) and enzyme concentrations (levels), we analyzed the capacity of different DoE fractional factorial designs and random sampling to find the enzyme concentration levels that optimize curcumin production. Each selected set of experiments was used to train a linear model containing main effects and two-factor interactions, and significant coefficients were determined by ANOVA. Enzymes with significant main effects are important for production, so, in the optimal strains, their expression levels should agree with the sign of their coefficients (high concentration for positive coefficients and low concentration for negative coefficients). For enzymes with insignificant main effects, production should not change regardless of the chosen concentration. Considering these criteria, we computed the frequency at which each strain is predicted as optimal by each design permutation or random sampling assuming 5% or 20% noise in the data. These predictions were compared to the "ground truth", defined by the curcumin production of the full factorial library obtained with the kinetic pathway model. We show here the results assuming 20% noise in the production data.

The full factorial data shows the presence of two strains with equal performance, characterized by high expression of C3H, CURS, and DCS, low expression of FCS, COMT, and CCOAOMT, and unaffected by the expression level of TAL (Figure 7.3A). Only resolution V and IV designs guarantee the identification of both or one of these two optimal strains (Figure 7.3A). However, while the resolution V design only suggests two strains as top producers, and the random selection of 64 strains results in the suggestion of four strains, the resolution IV design might suggest the construction of up to 16 new strains. Still, the targeted construction of the 16 strains required by the resolution IV design is more efficient than the random construction of 32 strains, as it always suggests one of the optimal strains, and the total number of strains to build and test is lower. When designs with lower resolution are chosen, the probability of finding the best strains from the full factorial library markedly decreases and, only in the case of the resolution III design, the number of suggested strains to construct is lower than in the random control (Figure 7.3A).

For the resolution IV design, we further studied whether including the best two producer strains in the design influenced the prediction of the top strains. In 60% of all the permuted resolution IV designs, the optimal strains were not included, which did not affect the predictions (Sup. Figure 7.3).

Considering that during *in vivo* studies experimental limitations might hinder the construction of some of the required strains for a design, we studied the robustness of resolution V and IV designs to missing strains. The performance of the resolution V design was minimally affected when up to 10 strains (16%) were excluded from the design. While the number of strains to construct in order to find the optimal production increased from 2 to 4, at least one of the two best producers was always suggested (Figure 7.3B). When the resolution IV design was used, excluding one strain from the design (6% of the library) had a minor impact on predictions. However, when 2 (13% of the library) or more strains were omitted, the probability of finding the best strain decreased, and the number of wrongly suggested optimal strains increased (Figure 7.3B).

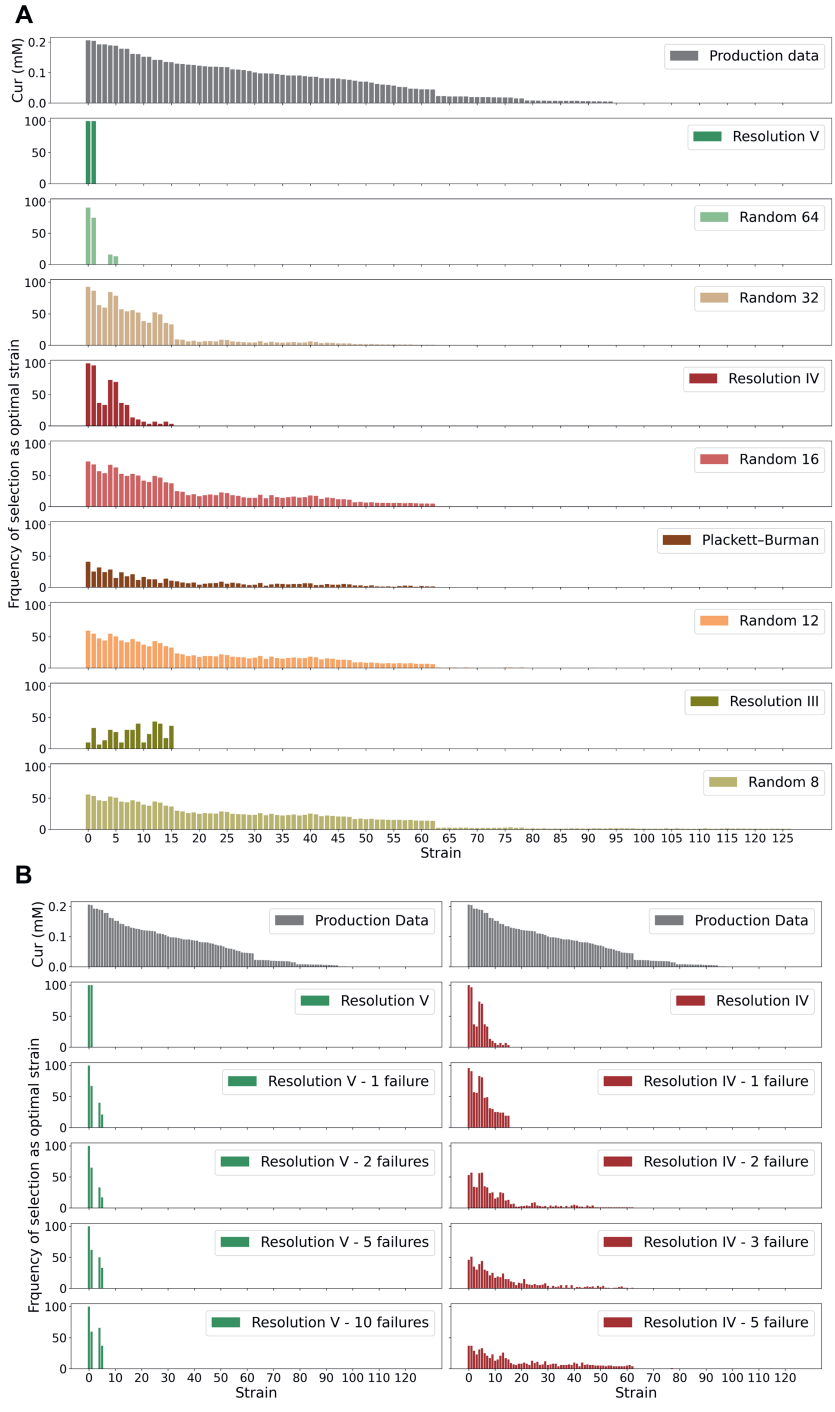


Figure 7.3: **A.** Prediction of optimal strains based on linear models trained with data from factorial designs and random sampling assuming 20% noise. **B.** Prediction of optimal strains by linear models trained with resolution V and IV factorial designs simulating the inability to construct some of the required strains assuming 20% noise.

Pathway insights: identification of significant factors and interactions

Besides the prediction of optimal strains, the analysis of the coefficients of the linear models aids the understanding of the studied pathway, unveiling the effect of each enzyme on curcumin production. These insights can then be used to guide following DBTL cycles focusing on the factors with the strongest influence on the response. Moreover, they can point to relevant interactions between factors that enhance the knowledge of the pathway. Here we assess the capacity of each of the DoE fractional factorial designs or random samples to identify significant main effects and interactions. The correct identification of these coefficients explains, in turn, the capacity of each design to find the optimal production strains.

The concentration of C3H is the factor with the strongest influence on production and, regardless of the level of noise, all the DoE designs identify C3H as a significant main effect with a positive influence on production (Figure 7.4A, B). The importance of this factor is also captured when 64 or 32 strains are randomly sampled. However, when 16, 12, or 8 strains are randomly selected, this effect is missed in 1%, 2%, and 7% of the experiments, respectively (Table 7.1).

While resolution V, IV, and III designs are always able to identify the negative effect of FCS and COMT, 6% to 20% of the PB designs, depending on the level of noise, are unable to capture this behavior (Figure 7.4A, B, Table 7.1). Similarly, the ability to identify the importance of these factors is lost when less than 64 strains are randomly sampled, especially when the level of noise increases.

Table 7.1: Frequency (expressed as a percentage) of main effects identified as significant by ANOVA using data from different fractional factorial designs or random strain sampling.

		C3H	FCS	COMT	CURS	CCOAOMT	DCS	TAL
Resolution V	5% noise	100	100	100	100	100	100	0
	20% noise	100	100	100	100	100	100	0
Random 64	5% noise	100	100	100	100	100	90.4	22.6
	20% noise	100	100	100	100	100	90.4	22.6
Random 32	5% noise	100	99.8	99.4	89.2	79.7	56.8	45.2
	20% noise	99.8	92.3	89.3	41.1	33.6	15.9	12.4
Resolution IV	5% noise	100	100	100	100	100	63.3	63.3
	20% noise	100	100	100	88.7	63.3	26.7	3.3
Random 16	5% noise	99.9	95.6	93.5	73.2	79.1	69.4	76.8
	20% noise	98.9	73.4	67.2	36	33.9	26.5	33.8
Plackett-Buramn	5% noise	100	95.7	93.6	93.6	89.8	90.1	87.9
	20% noise	100	79.2	77.6	70.1	65.8	69.4	59.4
Random 12	5% noise	99.7	94	90.2	72.8	77.6	68.9	82.9
	20% noise	97.7	69.1	65.6	40.1	38.6	33.7	46.8
Resolution III	5% noise	100	100	100	93.3	90	86.7	96.7
	20% noise	100	100	100	60	63.3	60	66.7
Random 8	5% noise	98	88.5	83.2	61.2	72.9	65.4	73.7
	20% noise	93	59.7	57.7	35.7	40.4	35.9	41.6

Given 5% noise, resolution V and IV designs, as well as the random selection of 64 strains, allow the identification of the positive and negative effects of CURS and COMT, respectively (Figure 7.4A, B, Table 7.1). However, when the level of noise increases, the chances of identifying these effects with resolution IV designs decrease to 89% and 63%, respectively. Yet, the likelihood of finding these effects is doubled with this design compared to the random selection of 32 strains. PB and resolution III designs show similar performance in identifying the importance of these genes, however, they are unable to correctly estimate their effect on the response as indicated by the high standard deviations of the coefficient values (Figure 7.4B). Notably, these designs show better performance than their random counterparts.

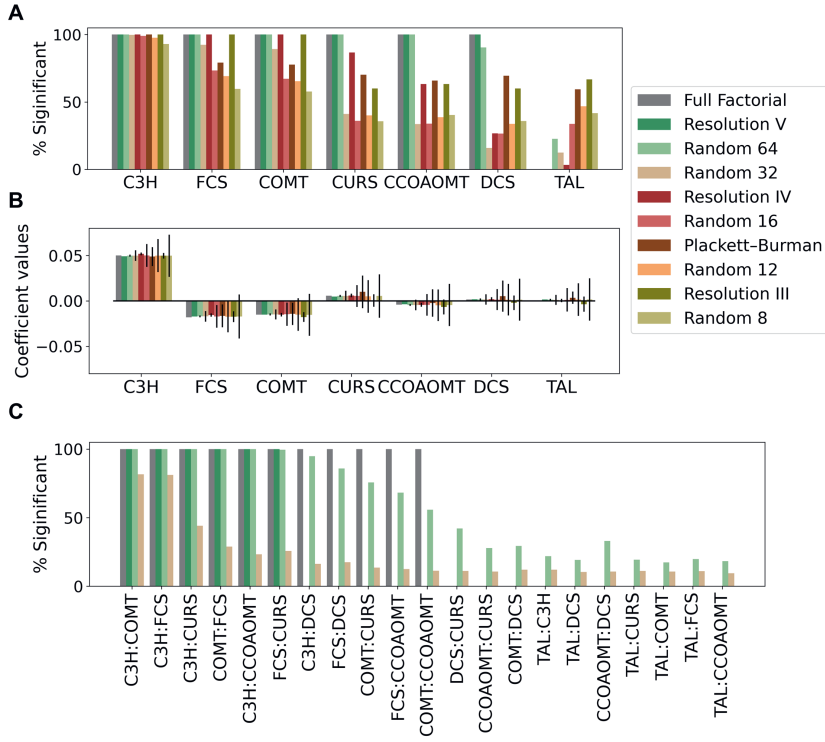


Figure 7.4: Estimation of linear model coefficients using data from fractional designs and random samples given 20% noise in the response **A.** Frequency of the identification of main effects as significant. **B.** Estimated coefficients for each main effect. Mean coefficients and standard deviations of all possible permutations of the design or random samples are shown. **C.** Frequency of the identification of 2-factor interactions as significant using data from full factorial, resolution V, random 64, and random 32 designs.

DCS and TAL are the factors with the smallest coefficients and, therefore, the smallest impact on production. However, while the expression level of DCS significantly affects production, modifying the expression of TAL does not change curcumin titers. The only design able to capture this behavior is the resolution V design, which outperforms the random selection of 64 strains. Other designs, based on DoE or a random selection of strains, are unable to distinguish the effect of these two genes (Figure 7.4, Sup. Table 7.1).

The inability to correctly identify significant main effects is the reason for the incorrect prediction of optimal strains by the designs (Figure 7.3A). For instance, models trained with resolution V designs only predict optimal strains with high concentrations of CURS and DCS, and low concentration of CCOAOMT. However, some of the linear models trained with other designs miss the relevance of these enzymes and incorrectly suggest strains with low CURS and DCS concentration and/or high CCOAOMT concentration as optimal.

In addition to the determination of main effects, resolution V designs or the random selection of 64 or 32 strains allow the estimation of all the coefficients corresponding to two-factor interactions (Figure 7.3C). These interactions point to factors whose effect on the response is affected by the level of another factor. When the full factorial data is used to train a linear model, thirteen significant two-factor interactions are found: all the enzymes but TAL have a significant interaction with C3H and FCS; CURS additionally interacts with COMT, DCS, and CCOAOMT; and COMT and CCOAOMT also show a significant interaction. The presence of a high number of significant interactions highlights the synergistic effect obtained when combining the optimal concentrations of various enzymes and underscores the relevance of combinatorial pathway optimization. However, not all the significant two-factor interactions have the same effect on the response and their absolute coefficients vary from $1.2 \cdot 10^{-2}$ to $6.2 \cdot 10^{-4}$ (Sup. Figure 7.4).

Assuming 5% noise in the response, linear models trained with resolution V designs correctly identify the eleven most important interactions, including interactions with absolute coefficients of 10^{-3}. When the level of noise increases to 20% this design still allows the identification of the six most important two-factor interactions, with absolute coefficients above $2.8 \cdot 10^{-2}$. Regardless of the level of noise, models trained with resolution V designs prevent the incorrect identification of insignificant interactions (false positives), frequently found when randomly selected strains are used for model training (Figure 7.3C).

When resolution IV designs are used to train linear models, specific two-factor interactions cannot be determined. However, the estimated coefficients of the confounded interactions give information on their relative importance compared to the main effects. Figure 7.5 shows how two-factor interactions 1 and 2 have an effect on the response similar to the main effect of COMT. Likewise, the effect of two-factor interaction 3 is similar to the main effect of CURS. Therefore, these designs are able to clearly identify that the effect of interactions in the studied system matters and should not be ignored. Notably, resolution IV designs with up to three missing strains are also able to correctly estimate the relevance of the two-factor interactions (Sup. Figure 7.5). Considering this, the best strains in the design space could be found using a sequential experimentation approach.

For instance, a resolution IV design could be first used to identify C3H, FCS, and COMT as the most important main effects, relative to the importance of two-factor interactions. In a second round, the expression level of these genes could be fixed according to the sign of their coefficients and a resolution V design with the remaining 4 factors could be performed. In this case, the resolution V design involves the construction of 16 strains and is equivalent to a full factorial design, which ensures the identification of the optimal strains with a total of 32 experiments.

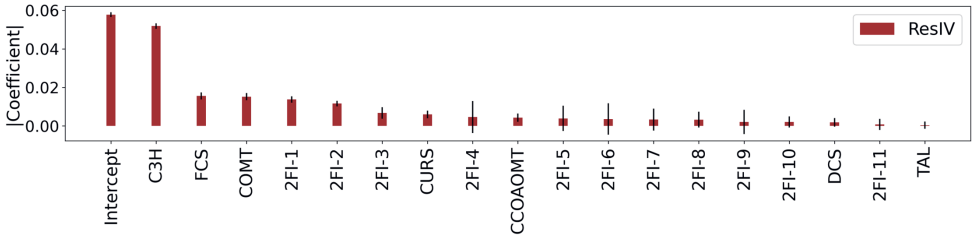


Figure 7.5: Absolute values of model coefficients trained with data from permutations of a resolution IV design. The mean coefficient and standard deviation of main effects and confounded two-factor interactions (2FI) are shown.

Finally, the partial correlation between main effects and interactions in PB designs should allow the identification of some interactions via subset regression to identify which factors and interactions result in models with better fit. However, subset models only consistently predicted the importance of C3H and failed to find significant interactions, showcasing the inappropriateness of this design for pathway understanding.

Factorial designs and machine learning

Machine learning (ML) algorithms can be used as an alternative to linear models to gather the information obtained during experimentation based on DoE designs [101]. Although recovering information from these models is harder than from linear models, ML algorithms can recognize more complex patterns within the data. As an example of a ML algorithm, we tested the ability of random forest models trained using 10-fold cross-validation or the complete datasets to predict the best-performing strains given data from random samples or DoE designs. However, training these models with 32 or fewer experiments often resulted in negative R^2 values for some of the iterations. Even when only models with R^2 coefficients above 0.6 were used, randomly selecting experiments or using DoE factorial designs resulted in equally bad predictions of best strains (Sup. Figure 7.6). Therefore, given the small number of experiments required when considering the optimization of seven factors, linear models outperform random forest. However, the ability of ML models to benefit from training data based on DoE designs if the number of factors and, therefore, experiments increases, remains unexplored.

Discussion

DoE involves the design of experiments using fractional factorial designs and their analysis using linear models and ANOVA. Here we showed how it can be used to find the optimal concentration of a pathway's enzymes and understand the impact of factors on production. In both cases, the resolution V design excels, providing the same information as the full factorial design and finding the strains with the best curcumin titers while requiring half of the experiments. When this design is used, all main effects are correctly identified as well as the most important two-factor interactions (Figure 7.4). We highlight the relevance of these interactions to understand and improve production, as designs where main effects and interactions are confounded struggle to find the optimal strains (Figure 7.3A). However, the identification of a significant interaction does not necessarily reflect a biological mechanism. For instance, factors representing all enzymes but TAL are included in significant interactions with C3H because only when C3H concentration is high, high levels of curcumin are obtained. This does not mean that all these enzymes physically interact with C3H but could lead to hypotheses aiming to explain the importance of this enzyme for the pathway functioning.

We propose resolution IV designs as the best trade-off between information gain and experimental effort, and the best option to initially screen the effect of factors in the response. The key strength of this design is the lack of confounding among main effects and interactions, which allows the confident identification of main effects. Besides, although two-factor interactions are confounded, these designs allow weighting their importance compared to the individual effects. Moreover, these designs provide a solid knowledge basis of the system under study that can be expanded in different directions depending on the experimental goal. Here, we show how, when the aim is to find the best possible production given the initial factors and levels, the most important main effects (compared to two-factor interactions) can be fixed and a resolution V design can be performed on the remaining factors. In this case, two-factor interaction coefficients involving the most important (fixed) factors will not be estimated, but the optimal strain will be found. As alternatives to the presented approach, different strategies can be pursued depending on the experimental goal. If the goal is to find coefficients for the most important factors, original resolution IV designs can be augmented using D-optimal designs. These designs select experiments from the full factorial that allow the clarification of the desired interactions by minimizing the variance of the model coefficients [72, 76]. Finally, when the researcher aims at expanding the original design space, the number of levels of the most relevant factors can increase following the direction indicated by the linear model coefficients. Alternative designs such as Box-Behnken designs that include three levels per factor can be used to train response surface models [72, 86], in this case, testing higher concentrations of C3H and lower concentrations of FCS and COMT.

In this study, the use of a kinetic model allowed the simulation of a full factorial design and the comparison of fractional designs without a limitation on throughput (*i.e.* number of strains to test and build). This comparison was performed considering realistic scenarios including noise and datasets with missing information due to, for instance, problems during strain construction. However,

during the *in vivo* optimization of pathways, the achievable throughput is a critical parameter that should determine how the optimization process is performed. Given the throughput, we recommend fixing the number of factors to screen to be able to obtain resolution IV designs. The advent of biofoundries that automate the strain construction process is continuously increasing the capacity to build strains [96, 338, 339]. This increase should be accompanied by high-throughput, automated cultivation and screening protocols as well as automated data collection [49]. Scaling these processes will allow the assessment of numerous factors in screening studies that should go beyond pathway engineering to include optimization at the metabolic and bio-process levels.

Declaration of interest

Joep Schmitz is employed by dsm-frimenich.

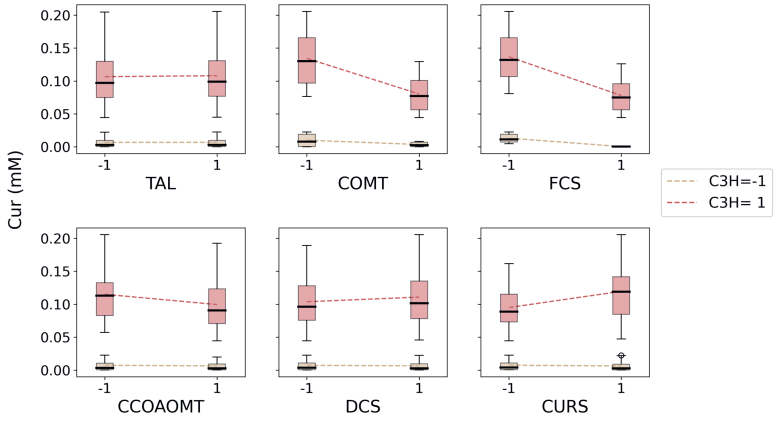
Acknowledgment

This project was founded by NWO (project number GSGT.2019.008).

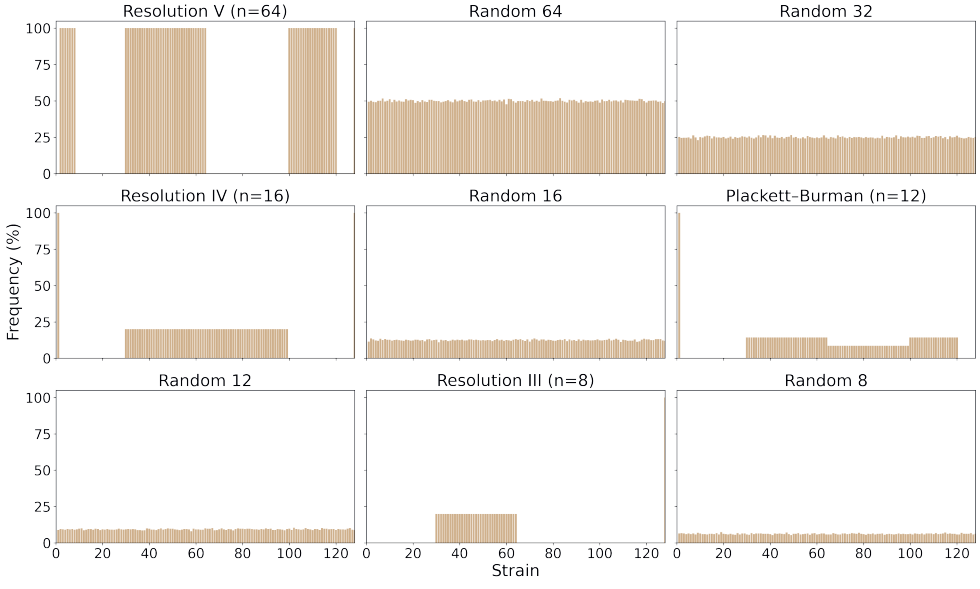
Data availability

Scripts for the generation and analysis of the designed libraries, the prediction of optimal strains and the analysis of the desigs using ML, as well as additional data are available at Gitlab.

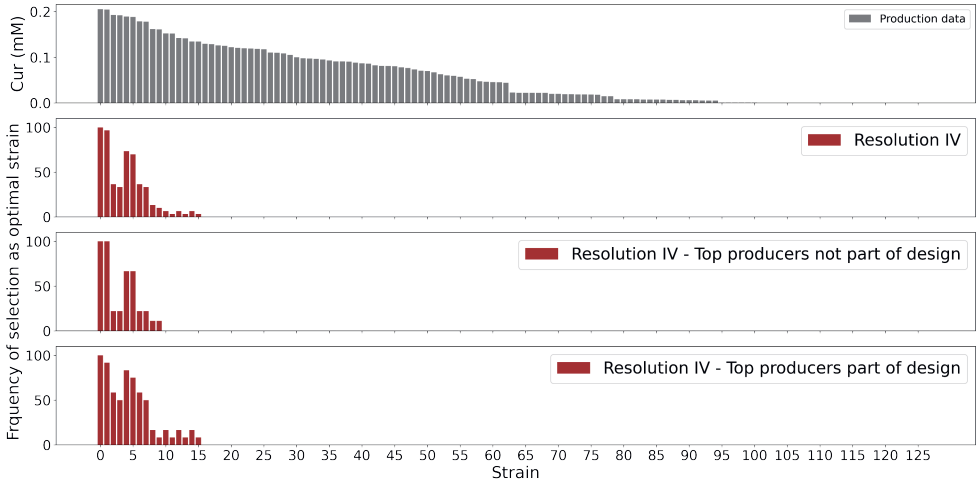
Supplementary Figures



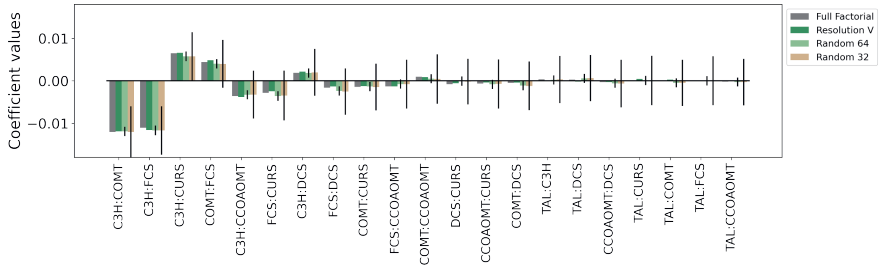
Sup. Figure 7.1: Sequential vs. combinatorial experimentation. Example of how the concentration of the C3H enzyme affects the impact of changing other enzyme concentrations on curcumin (Cur) production. -1, low enzyme concentration; 1, high enzyme concentration.



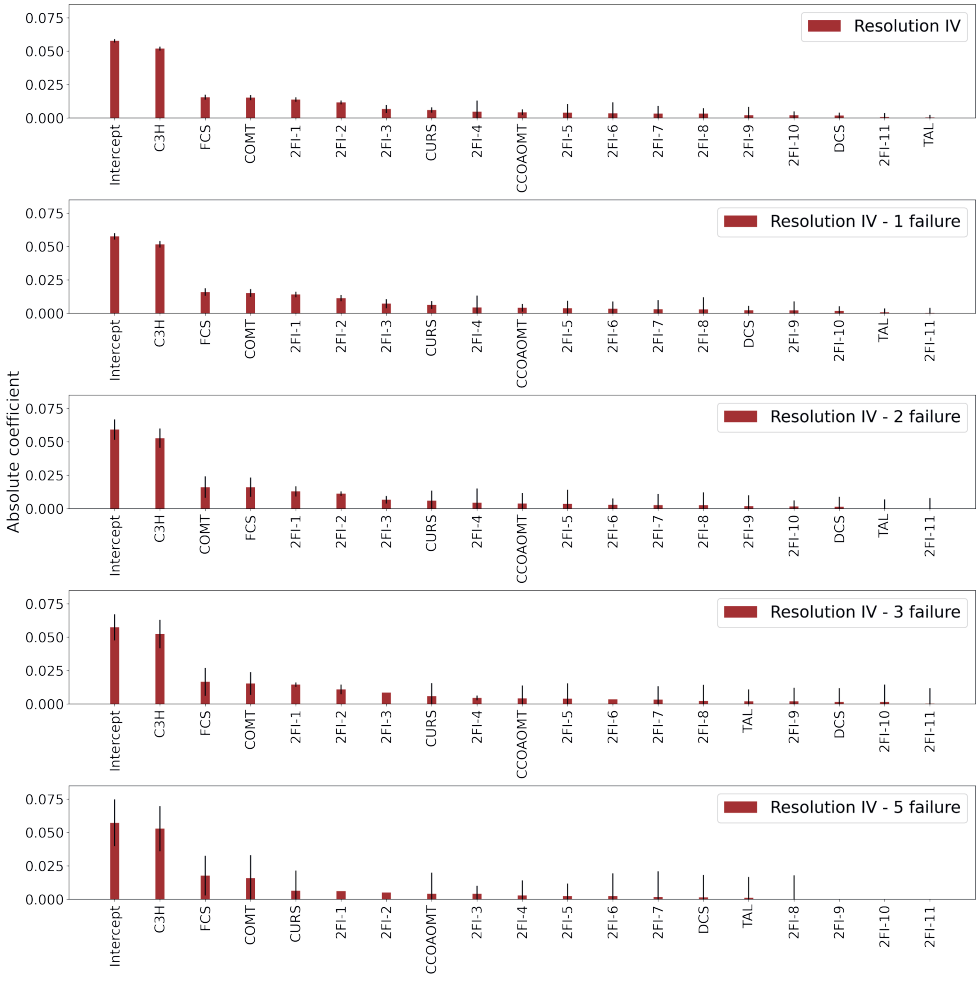
Sup. Figure 7.2: Frequency of strain selection by the different design approaches including factorial designs and random sampling.



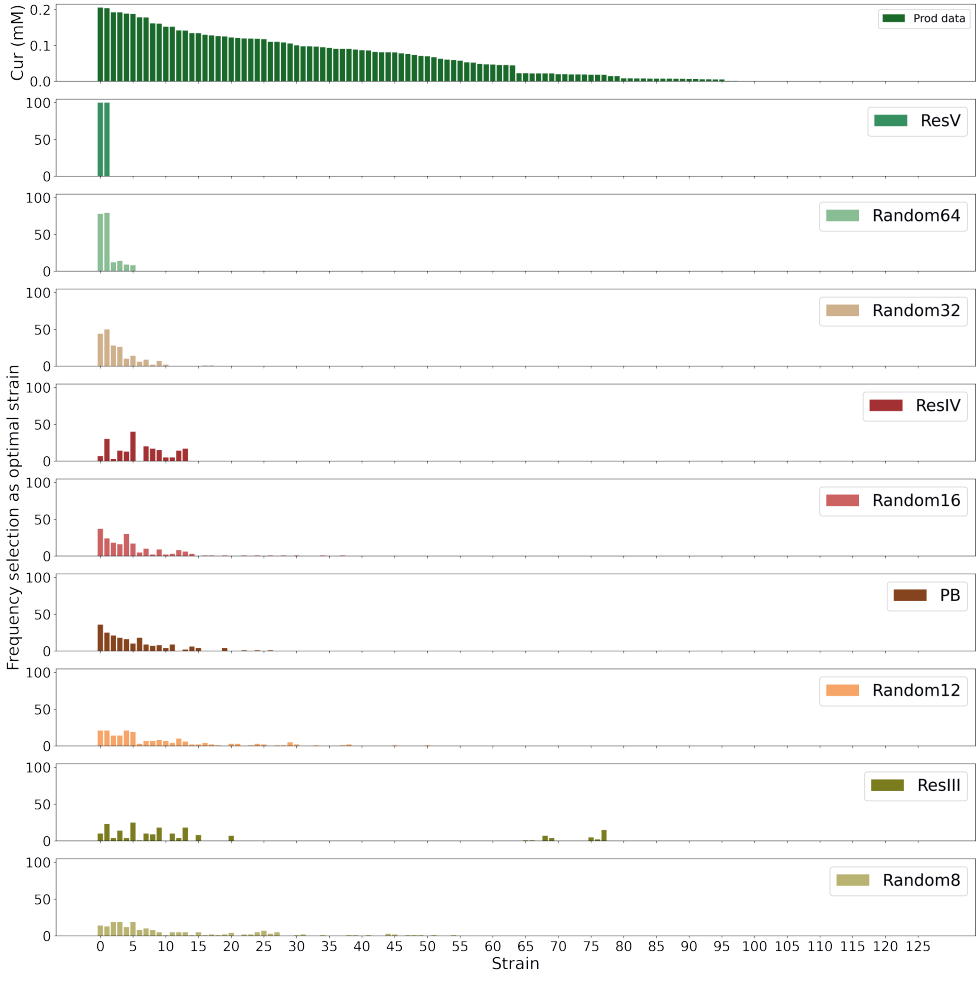
Sup. Figure 7.3: Prediction of best strains by linear models trained with all resolution IV designs, resolution IV designs that exclude the two best strains in the design (60% of the designs) or that include them (40% of the designs).



Sup. Figure 7.4: Estimated coefficients for each two-factor interaction using data from resolution V, random 64 and random 32 designs. 20% noise in the response is used. The mean coefficient and standard deviation of the coefficients considering all possible permutations of the design or random samples are shown.



Sup. Figure 7.5: Absolute coefficients of the main effects and the 11 confounded two-factor interactions (2FI) estimated using data from resolution IV designs with missing strains. The mean coefficient and standard deviation of the coefficients considering all possible permutations of the design are shown.



Sup. Figure 7.6: Prediction of best strains by random forest models trained with data from different factorial designs or random selection of strains assuming 20% noise in the response.

CHAPTER 8

Combinatorial optimization of pathway, process and media for the production of p-coumaric acid by *Saccharomyces cerevisiae*

Sara Moreno Paz*, Rianne van der Hoek*, Elif Eliana, Vitor A. P. Martins dos Santos, Joep Schmitz#, María Suárez Diez#

*Contributed equally, #Jointly supervised this work

This chapter is published in *Microbial Biotechnology*
doi.org/10.1111/1751-7915.14424

Abstract

Microbial cell factories are instrumental in transitioning towards a sustainable bio-based economy, offering alternatives to conventional chemical processes. However, fulfilling their potential requires simultaneous screening for optimal media composition, process, and genetic factors, acknowledging the complex interplay between the organism's genotype and its environment. This study employs statistical Design of Experiments (DoE) to systematically explore these relationships and optimize the production of p-coumaric acid (pCA) in *Saccharomyces cerevisiae*. Two rounds of fractional factorial designs were used to identify factors with a significant effect on pCA production, which resulted in a 168-fold variation in pCA titer. Moreover, a significant interaction between the culture temperature and expression of ARO4 highlighted the importance of simultaneous process and strain optimization. The presented approach leverages the strengths of experimental design and statistical analysis and could be systematically applied during strain and bio-process design efforts to unlock the full potential of microbial cell factories.

Introduction

Microbial cell factories play a pivotal role in driving the transition towards a bio-based economy, being a sustainable alternative to traditional chemical processes [24]. Microorganisms can efficiently transform raw materials into valuable products such as bulk chemicals or pharmaceuticals. However, to unlock their potential for biotransformation in an economically feasible manner, it is essential to optimize production pathways and bio-processes [243].

Pathways can be optimized sequentially by tuning individual genetic factors in isolation. However, this does not capture the complex interplay between different genetic elements and the products they code for. It is hence desirable to perform combinatorial pathway optimization, which is based on the simultaneous optimization of multiple genetic factors and facilitates the identification of complex interactions [73, 332]. Moreover, the overall performance of the microbial cell factory is not only determined by its genotype but is also influenced by the production conditions, as factors such as media nutrients, pH, cultivation temperature, and aeration influence cell physiology and metabolism. Strains are usually optimized holding the environmental conditions constant and only the most promising strain advances to the bio-process optimization stage [32, 85]. However, this approach might ignore genetic designs that, although inferior in standard laboratory conditions, have bigger potential when the media and bio-process are optimized [88]. Only by simultaneously screening for optimal media composition and genetic factors, the dynamic interplay between the organism's genotype and the environment in which it operates can be considered [88, 89].

Combinatorial optimization of strains, media, and process parameters, however, requires exponentially increasing resources. Statistical design of experiments (DoE) allows a structured exploration of the relationships between experimental variables (factors) and the measured response. Full factorial designs are a type of DoE design that tests all possible combinations of factor levels, characterizing factor effects and allowing the estimation of interactions. The number of experiments to be performed depends on the number of genetic and environmental factors to be tested (*e.g.* expression of a gene, temperature) and the number of levels per factor (*e.g.* low, medium and strong gene expression, 20°C, 25°C, 20°C, and 35°C) according to $\prod_{i=1}^{F} \mathrm{L}_i$, where F is the number of factors and L_i is the number of levels of factor i. In these designs, the effect of a factor is not only estimated considering replicate experiments but also all the experiments where the given factor is constant regardless of other factor levels. This property can be leveraged in fractional factorial designs that reduce the number of experiments to perform while maximizing the information gain. This is achieved by performing experiments that preserve orthogonality in the desired factors, *i.e.* ensuring that the effect of a factor is not confounded by planned changes in other factors. The generated data is fitted to a linear model so main effects (MEs), representing the impact of not-confounded factors on the response, are identified. Similarly, the so-called two Factor Interactions (2FI) that occur when the effect of a factor on the response changes based on the level of another factor, can also be estimated [72].

Although decreasing the number of experiments ensures the identification of not-confounded effects, information regarding confounded factors or interactions is lost [72]. For example, resolution IV designs, a type of fractional design, allows the identification of main effects but confound 2-factor interactions among each other. Therefore, these designs can be used to report if 2-factor interactions are important but cannot clarify which 2-factor interactions have a significant effect on the response. Fractional designs with lower resolution, such as resolution III designs, require fewer experiments but confound main effects with 2-factor interactions. These designs should therefore only be used when interactions among factors are not expected, rarely true for biological systems. Alternatively, designs with higher resolution ensure the identification of interactions at the expense of a higher experimental workload.

We used production of p-coumaric acid (pCA) by *Saccharomyces cerevisiae* as an example of DoE-aided combinatorial pathway, media, and process optimization. pCA can be produced from phenylalanine (Phe), an aromatic amino acid produced within the shikimate pathway. It is a precursor for a wide array of biologically relevant molecules such as pharmaceuticals, flavors, fragrances, and cosmetics [340]. Although pCA production has been independently optimized at the strain and bio-process levels [340, 341, 342, 343, 344], we show the interplay between genetic and environmental factors highlighting the importance of simultaneous process and strain optimization.

Material and methods

Strain construction

Promoter, terminator, and ORFs sequences from *aro4*, *aroL*, *aro7*, *pal1*, *c4h*, and *cpr* codon optimized for *S. cerevisiae* were obtained from Moreno-Paz et al. [345] (Sup. Table 1). Cassettes formed by combinations of promoter, ORF, and terminator (Sup. Table 1) were assembled via Golden Gate into a backbone plasmid containing a 50 bp homologous connector sequence to facilitate *in vivo* recombination of the gene cluster [346]. Golden Gate products were transformed into competent *Escherichia coli* DH10B cells, plasmids were isolated, and cassettes were confirmed by PCR.

Strains were constructed as described in Moreno-Paz et al. [345]. In short, a host strain with Cas9 integrated in the non-coding region between YOR071c and YOR070c in chromosome 15 was transformed with a linear guide RNA targeting the AEHG01000256.1 locus (210ng/kb) [347], equimolar cassettes for the required designs (Table 8.1, Sup. Table 2) (100-300ng/kb) and linear backbone fragments (35ng/kb) following the LiAc/ssDNA/PEG method [348]. The connector sequences on the cassettes facilitate *in vivo* recombination of a cluster of genes in the genome [346]. Transformants were plated on Qtray (NUNC) with 48-divider (Genetix) containing YEPhD agar medium and selection agent. Colonies appeared on the plate after 3 days of incubation at 30°C. Single colonies were picked with Qpix 420 (Molecular Devices) into 96 well plates containing YEPhD agar medium and selection agent and regrown for 3 days at 30°C. Colonies were confirmed using whole genome sequencing and correct strains were stored at -80°C.

Table 8.1: Structure of the gene clusters. Cell values indicate the promoter used for each gene. Promoters were selected from Moreno-Paz et al. [345].

Gene cluster	ARO4	AROL	ARO7	PAL1	C4H	CPR1
1	TDH3	TEF1	ACT1	RPS9A	CHO1	CCW12
2	TDH3	TEF1	ACT1	VMA6	PXR1	CCW12
3	TDH3	TEF1	PFY1	RPS9A	CHO1	CCW12
4	TDH3	TEF1	PFY1	VMA6	PXR1	CCW12
5	TDH3	RPL28	ACT1	RPS9A	CHO1	CCW12
6	TDH3	RPL28	ACT1	VMA6	PXR1	CCW12
7	TDH3	RPL28	PFY1	RPS9A	CHO1	CCW12
8	TDH3	RPL28	PFY1	VMA6	PXR1	CCW12
9	MYO4	TEF1	ACT1	RPS9A	CHO1	CCW12
10	MYO4	TEF1	ACT1	VMA6	PXR1	CCW12
11	MYO4	TEF1	PFY1	RPS9A	CHO1	CCW12
12	MYO4	TEF1	PFY1	VMA6	PXR1	CCW12
13	MYO4	RPL28	ACT1	RPS9A	CHO1	CCW12
14	MYO4	RPL28	ACT1	VMA6	PXR1	CCW12
15	MYO4	RPL28	PFY1	RPS9A	CHO1	CCW12
16	MYO4	RPL28	PFY1	VMA6	PXR1	CCW12

pCA production experiments

Single colonies were grown in 10 ml YPDA media (Takara) in 50 ml tubes for 24h. Cultures were washed and inoculated in minimal media at starting OD of 0.3 or 0.6 according to the experimental design. Minimal media contained 20 g/l glucose (Acros Organics), and 1.7 g/l yeast nitrogen base without amino acids or ammonium sulfate (BD Difco). A 60.5 mM nitrogen concentration in the media was obtained with 4 g/l ammonium sulfate (Acros Organics) or 1.82 g/l urea (Acros Organics). When required, media was buffered at a pH of 7 using 126 mM Na_2HPO_4 (Acros Organics) and 18 mM citric acid (Sigma-Aldrich) [349] and/or supplemented with 5 mM phenylalanine (Sigma-Aldrich) and/or 5 mM glutamic acid (Sigma-Aldrich). Cells were grown for 48 h at the required temperature and agitation speed in 50 ml mini-bioreactor tubes (Corning) in an Innova 44 incubator (New Brunswick Scientific). At the end of the cultivation samples for OD measurements and pCA quantification were taken.

pCA quantification

For pCA quantification, 400 µl of culture were mixed with 800 µl of acetonitrile (Thermo Scientific) and centrifuged for 10 min at 4000 g. The acetonitrile phase was used for analysis using high performance liquid chromatography (HPLC) on a Shimadzu LC2030C Plus 2 machine equipped

with a Poroshell 120EC-C18 column (250 x 4.6 mm, Agilent) and a UV/vis detector. Mobile phase was used at a rate of 1 ml/min and was composed of Milli-Q water (A), 100 mM formic acid (B), and acetonitrile (C) at varying proportions: 77:10:13 (v/v/v) in the first 10 min, 23:10:67 (v/v/v) in the next 9 min, and 77:10:13 (v/v/v) in the last six minutes. pCA was detected at a wavelength of 280 nm. Standards were prepared using pCA purchased from Sigma-Aldrich.

Experimental design and statistical analysis

The FrF2 function from the FrF2 R package was used for the generation of the designs given the number of factors and the desired resolution [337].

Experimental data was used to train a linear model:

$$y = \beta_0 + \sum_{i=1}^{i=n} ME_i \cdot F_i + \sum_{i=1}^{i=n} \sum_{j=1}^{j=n} 2FI_{i:j} \cdot F_i \cdot F_j, \tag{8.1}$$

where y represents the pCA concentration; β_0 refers to the y-intercept, ME_i represents the main effect of factor i (F_i) and $2FI_{i:j}$ refers to the two-factor interaction between factor i and j. The total number of factors is indicated by n.

Ordinary least squares regression minimizing the sum of squared differences between the observed and predicted values was used to estimate the coefficients for each term in the model (ME_i, $2FI_{i:j}$) using the R lm function. Then the summary function was used to obtain the ANOVA table which provides the estimated coefficients and their associated p-values. p-Values were corrected using Bonferroni. The adjusted coefficient of determination (Adj R^2) and the mean absolute error (MAE) were used to assess the model fit to experimental data.

Results

Selection of genetic and environmental factors and levels

The shikimate pathway is tightly regulated and aromatic amino acids exert feedback inhibition on some of its enzymes (Figure 8.1) [350]. Expression of feedback-resistant variants of ARO4 ($ARO4^{K229L}$) and ARO7 ($ARO7^{G141S}$) are common strategies to increase pCA production [340, 342]. Besides, the phosphorylation of shikimate performed by ARO1 has been hypothesized as rate-limiting step in the pathway. Rodriguez et al. reported a beneficial effect of expressing *E. coli* AROL to increase the flux through this reaction [342]. Therefore, we selected the expression of $ARO4^{K229L}$, $ARO7^{G141S}$, and AROL as genetic factors with the potential to affect pCA production. Each of these factors was evaluated at two levels based on the strength of the promoter-terminator pair assigned to each gene (Table 8.2) [345].

To produce pCA from Phe, the expression of two heterologous genes is required: phenylalanine ammonia lyase (PAL) and cinnamate 4-hydroxylase (C4H). Although *S. cerevisiae* contains endogenous cytochrome P450 reductases (CPR), expression of a C4H-associated CPR is recommended [340, 341, 342, 343]. We expressed *Arabidopsis thaliana* CPR under a constitutive promoter

and considered the expression of PAL and C4H as an additional genetic factor for the design. The expression levels of PAL and C4H were evaluated using two promoter-terminator pairs (Table 8.2).

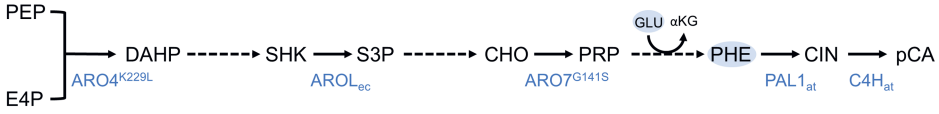


Figure 8.1: p-Coumaric acid (pCA) production pathway. Genes whose expression is considered as factor for the design are shown. The origin of the gene is indicated as a subscript: ec, *E. coli*; at, *A. thaliana*. Glutamic acid (Glu) and phenylalanine (Phe) are highlighted as they are selected as factors for media optimization. PEP, phosphoenolpyruvate; E4P, erithrose-4-phosphate; DAHP, 3-deoxy-7-phosphoheptulonate; SHK, shikimate; S3P, shikimate-3-phosphate; CHO, chorismate; PRP, phrephenate; PHE, phenylalanine, CIN, cinnamate; GLU, glutamate; αKG, α-ketoglutarate.

Temperature (T), agitation (rpm), and initial cell density (OD) are usual variables tuned during bio-process optimization and were selected as factors to improve pCA titers [80, 81, 82, 141, 351, 352]. Temperature was varied between 30°C, the optimal growth temperature of *S. cerevisiae*, and 20°C, as lower temperatures might improve heterologous pathway expression [351]. Agitation and initial OD were varied between 180-250 rpm and 0.3-0.6, respectively (Table 8.2).

Table 8.2: Factors and levels used for pCA optimization.

Factor	Low-level (-1)	High-level (1)
Temperature	20 °C	30 °C
Agitation	180 rpm	250 rpm
Initial cell desnity	0.3	0.6
N-source	Urea	Ammonium sulfate
pH	Unbuffered	Buffered
Phe	0	5 mM
Glu	0	5 mM
PAL-C4H promoters	VMA6 - PXR1	RPS9A - CHO1
ARO7 promoter	PFY1	ACT1
AROL promoter	RPL28	TEF1
ARO4 promoter	MYO4	TDH3

Combes et al. showed that the pH of the media affects pCA production [353]. Acidic pH, below the pKA of pCA (4.65), favors the undissociated form of pCA (pHCA) in the media that diffuses into the cell where it dissociates (pCA^-), acidifying the cytoplasm and requiring active export at the cost of ATP. Considering this, two factors that influence the pH of the media were selected: the addition of a buffer and the use of different nitrogen sources (Table 8.2). When ammonium sulfate or urea are used as nitrogen source, pH below and above the pKA of pCA are

expected respectively [349]. Independently of the N-source used, the citrate phosphate buffer can control the pH of the culture but can negatively impact cell growth [349].

Media supplementation is an additional common strategy to increase production [79, 83, 85, 141]. Phenylalanine is the substrate of PAL, the first enzyme required for the production of pCA and glutamate is the nitrogen donor used during Phe production (Figure 8.1). Therefore, the additions of these amino acids were considered as additional factors (Table 8.2).

Resolution IV design: impact of individual factors on pCA production

The effect of changing process conditions and media-related factors on pCA production was evaluated using a resolution IV fractional factorial design. This design allows the estimation of main effects of all the factors while confounding 2-factor interactions. They can be used during screening to identify factors with a significant impact on production that can be the focus of later optimization.

In order to obtain a resolution IV design with 11 factors (Table 8.2), 32 experiments are required (Sup. Table 3). Although traditional applications of DoE use single-replicate screening, replicates are a necessity to assess biological variation and, for each experiment, pCA production was measured in three independent cultures [73]. These experiments involved the construction of 16 strains including all possible combinations of the four selected genetic factors. Each strain is then tested in two different conditions determined by the design. However, strains containing gene clusters 2 and 6 (Table 8.1) could not be constructed and the effect of not performing four out of the 32 experiments was evaluated. Reducing the number of experiments to 28 did not affect the estimation of main effects but increased the complexity of the confounding patterns for the 2-factor interactions. Considering the goal of the experiment was to determine the main effects, construction of the two additional strains was not required, which accelerated the implementation of the design round. In the 28 experiments performed, pCA production varied two orders of magnitude, from 1.3 mg/l to 158.2 mg/l, confirming the impact of the selected factors on production (Figure 8.2A).

Linear models containing only main effects or main effects and confounded 2-factor interactions were trained. Including 2-factor interactions increased the coefficient of determination from 0.66 to 0.94 (Figure 8.2C). Figure 8.2B shows the estimated coefficients of the model including main effects and confounded 2-factor interactions. An ANOVA was used to determine the significance of each main effect and 2-factor interaction on pCA production and p-values were corrected using Bonferroni. All factors but T, OD, Glu, and ARO7 had a significant effect on pCA production. Although, seven of the estimated 2-factor interactions were also significant, identifying the specific significant 2-factor interactions was not possible due to their confounding.

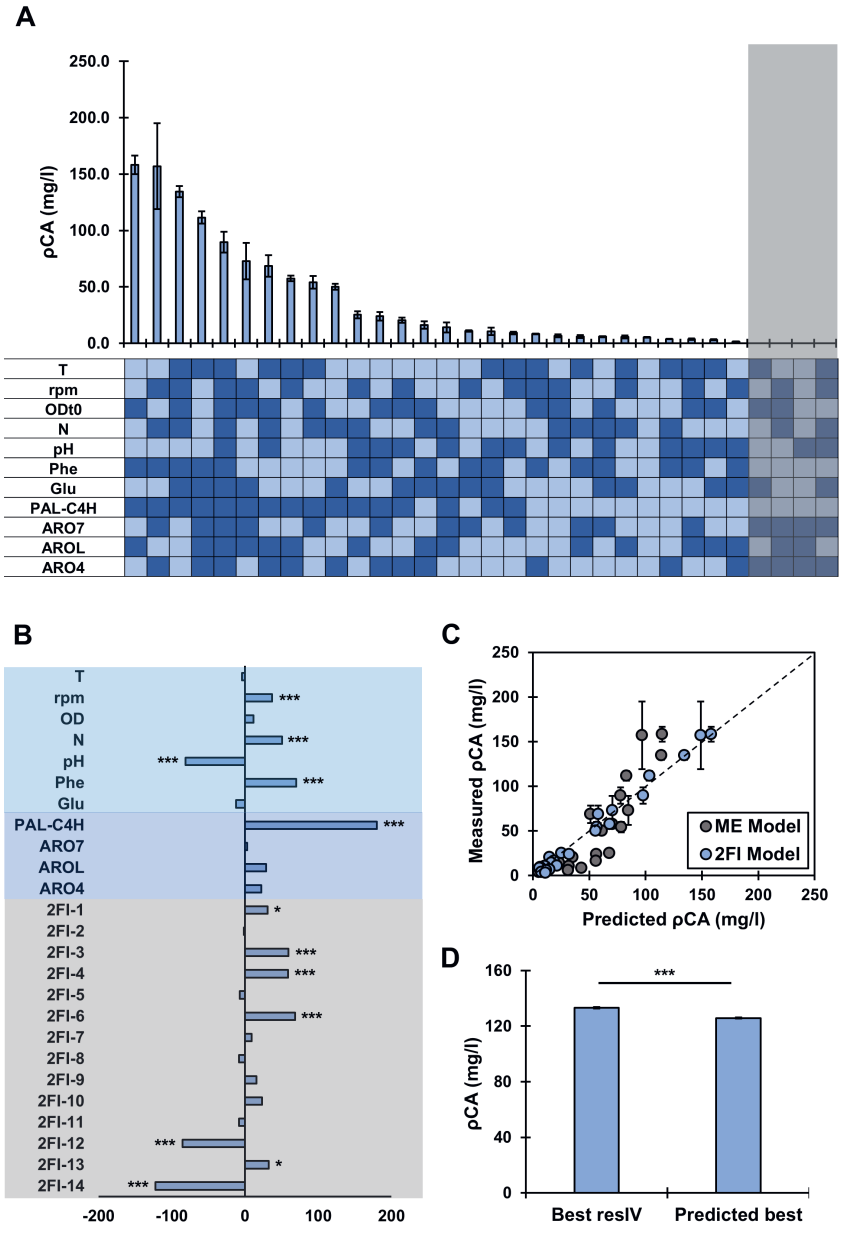


Figure 8.2: Resolution IV Design **A.** Measured pCA production at different combinations of factors and levels. Light colors indicate low level (-1) and dark blue indicates high level (1). The shaded gray area indicates the four experiments that could not be performed. **B.** Coefficients of the model including main effects and confounded 2-factor interactions (2FI). ** indicates corrected p-value ≤ 0.001, ** ≤ 0.01 and * ≤ 0.05. **C.** Fit of models including main effects (ME) or main effects and two-factor interactions (2FI model) to experimental data. ME model: Adj R^2=0.66, MAE = 129; 2FI model: Adj R^2=0.94, MAE = 33. **D.** pCA production in validation experiment.

The main effect with the highest impact on performance was the expression strength of PAL and C4H, with a positive regression coefficient. This indicates that a high expression of the heterologous genes for pCA production is essential to obtain high titers. The effect of PAL-C4H was followed by the negative impact of buffering of the media represented by a negative regression coefficient. Although the addition of a buffer could control the dissociation of pCA [353], it negatively affected cell growth and resulted in overall low pCA titers. The third most relevant main effect was the addition of Phe, with a positive coefficient that shows the benefit of Phe supplementation on pCA production [354].

Notably, estimated coefficients for 2FI-6, 12, and 14 had a similar impact on pCA titer than PAL-C4H, pH, and Phe, and the coefficient of determination was significantly improved when 2-factor interactions were considered for pCA production (Figure 8.2B, C), indicating that a design with a higher resolution that allows the estimation of 2-factor interactions is required to optimize pCA production. The importance of 2-factor interactions was further confirmed in an independent experiment where the best experiment from the resolution IV design was compared to the best predicted experiment according to the model's main effects. The predicted best experiment showed a small (5.6%) but significant reduction in pCA production, confirming the importance of 2-factor interactions (Figure 8.2D, Sup. Table 4).

Resolution V design: identification of relevant 2-factor interactions

Fractional factorial resolution V designs are required to identify main effects and 2-factor interactions. When 11 factors are considered this design includes 128 experiments. In order to decrease the number of experiments, factors with the highest impact on pCA production were fixed: only strains with high expression of PAL-C4H were considered and unbuffered media supplemented with Phe was used. This reduced the number of factors to eight and the number of required experiments to 64 (Sup. Table 5). Considering the small variation of the resolution IV dataset, duplicates instead of triplicates were used during this round. The top producing conditions from the resolution IV experiments was included as control.

In the resolution V experiments, pCA production varied from 79.8 mg/l to 218.7 mg/l, improving the maximum production found in the first round by 38% (Figure 8.3A, Sup. Table 5). The use of strains with high expression of PAL-C4H in unbuffered media supplemented with Phe, increased the minimum production of pCA in this round by 62%, supporting the information provided by the resolution IV model.

Experimental data was used to train new linear models based on main effects or main effects and 2-factor interactions. The model trained with main effects showed a coefficient of determination of 0.55 that increased to 0.74 when 2-factor interactions were considered, highlighting the relevance of 2-factor interactions to explain pCA production (Figure 8.3C).

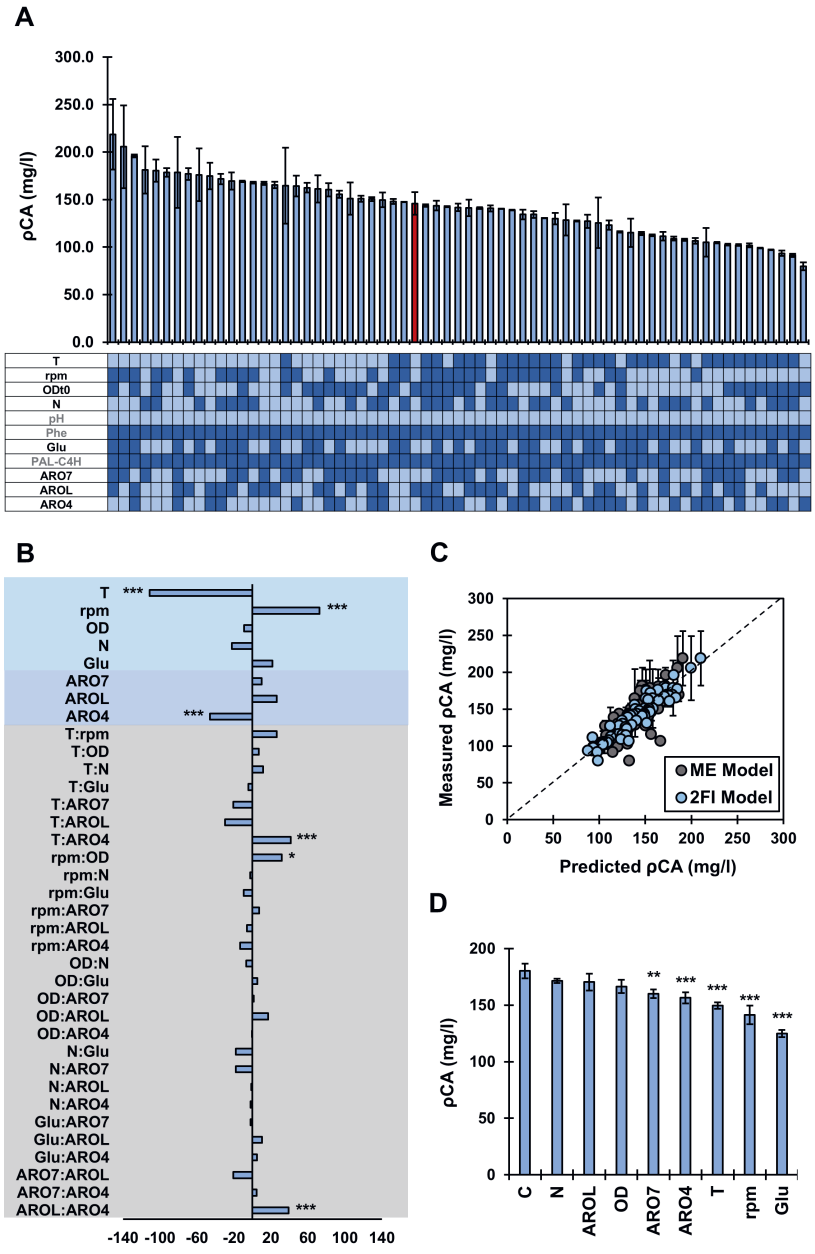


Figure 8.3: Resolution V Design **A.** Measured pCA production at different combinations of factors and levels. Light blue indicates low level (-1) and dark blue indicates high level (1). The red bar is the best performing experiment from the resolution IV round. Factors in grey were fixed based on inforation from the resolution IV design. **B.** Coefficients of the model including main effects and 2-factor interactions. *** indicates corrected p-value ≤ 0.001, ** ≤ 0.01 and * ≤ 0.05. **C.** Fit of models including main effects (ME) or main effects and two-factor interactions (2FI model) to experimental data. ME model: Adj R^2=0.55, MAE = 82; 2FI model: Adj R^2=0.74, MAE = 51. **D.** pCA production in validation experiment.

Temperature and agitation were identified as significant process-related factors, so low T and high agitation (rpm) improve pCA production (Figure 8.3B). ARO4 was the only significant genetic factor, and, in contrast to other reports, low expression of this gene positively affected pCA titers (Figure 8.3B) [340, 342]. Moreover, three significant positive 2-factor interactions were found: T:ARO4, rpm:OD, and AROL:ARO4 (Figure 8.3B).

In order to find the optimal strain and conditions for pCA production, the model including main effects and 2-factor interactions was used to predict pCA titers for all strains in all possible media conditions. The use of a strain with high expression of PAL-C4H, ARO7, and AROL and lower expression of ARO4 in a media supplemented with urea, Phe, and Glu incubated at 20°C and 250 rpm with an initial OD of 0.3 was predicted to optimize pCA production. These conditions were met by the top producing experiment measured in the resolution V round. To avoid the bias toward performed experiments during the estimation of model parameters, a new model was trained excluding data from the top producing experiment. When pCA production was predicted, the excluded top producer experiment was still suggested as optimal. Moreover, we evaluated the effect of individually changing each factor to its sub-optimal level. As expected, these modifications decreased or did not affect pCA production (Figure 8.3D, Sup. Table 6). Changing the initial OD, the nitrogen source, or the expression of AROL - all factors with no significant main effects - did not significantly change the pCA produced. In contrast, modifying the expression of ARO4, T, and rpm, factors with significant main effects, negatively impacted pCA production. Although main effects related to ARO7 and Glu were insignificant, reducing the expression of ARO7 and omitting Glu supplementation, negatively impacted pCA production, which could be explained by higher-order factor interactions not included in the model.

Interaction graphs were used to understand the relationship between factors involved in significant 2-factor interactions and pCA production. Figure 8.4A shows the interaction between a genetic factor, ARO4, and a process-related factor, temperature. While at 30°C expression of ARO4 does not affect production, a lower expression results in a higher titer at 20°C. Genetic factors also interact with each other, as an imbalanced expression of low AROL and high ARO4 results in the lowest pCA titer (Figure 8.4B). Last, a significant interaction between rpm and OD was found since, although fast agitation is always preferred, it has a higher impact in cultures with high initial cell density (Figure 8.4C).

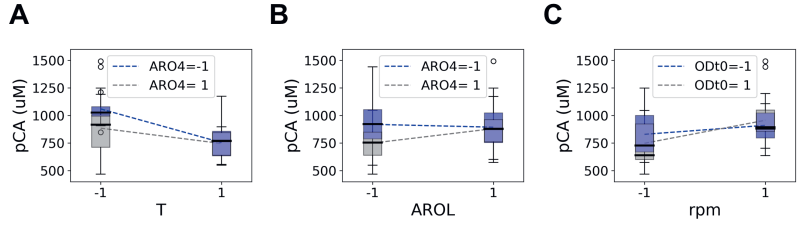


Figure 8.4: Interaction plots of significant 2-factor interactions. See Table 8.2 to identify levels corresponding to -1 and 1. T, temperature; rpm, agitation speed; ODt0, initial cell density.

Discussion

DoE has been commonly applied to process optimization [79, 80, 81, 82, 83, 84, 85] and strain design [32, 86, 87]. However, only a few studies consider the simultaneous optimization of genetic and environmental factors [88, 89]. Importantly, these studies showed that the interplay between both types of factors must be taken into account simultaneously for the optimization of microbial conversion processes. Our work underscores and strengthens these findings, using pCA production in *S. cerevisiae* as an example. The importance of simultaneous process and strain optimization is highlighted by the T:ARO4 interaction (Figure 8.4A). If strain selection had been performed at *S. cerevisiae*'s standard growth temperature (30°C), tuning the expression of ARO4 would have been considered irrelevant. However, given that the expression of ARO4 becomes important at lower temperatures, a sub-optimal strain could have been selected for subsequent process optimization.

The interplay between the strain performance and the bio-process design is especially important when moving from laboratory-scale to large-scale processes. This step-wise endeavor is time-consuming, labor-intensive and expensive. It thus benefits from scale-down experimentation [355]. The central paradigm of scaling-down states that scale-up will succeed when changes in the cellular environment caused by changes in scale do not influence cell behavior [356]. Here we show how DoE can identify genetic and process parameters with significant influence on production that should be the focus of the down/up-scaling plan.

We used DoE to understand the effect of seven process-related factors (T, rpm, OD, N, pH, Phe, and Glu), four genetic factors (PAL-C4H, ARO7, AROL, and ARO4), and their interactions on pCA production. Considering two levels per factor, 2048 experiments would be required to find all possible interactions between factors and ensure the identification of the best production conditions. Instead, we performed 92 experiments (4.5% of the total) divided into two consecutive rounds. The first round identified the expression of PAL-C4H, the addition of Phe, and the use of unbuffered media as key variables to ensure high pCA titers. These findings were taken up in a second experimental round for which the optimal pCA production conditions were found: incubation at low temperature and high agitation of a strain with low expression of ARO4. A 38% increased pCA production was obtained in the resolution V round compared to the best experiment in the resolution IV design. Moreover, a 168-fold variation was measured between the worst and best performed experiments.

As indicated, only fourteen of the sixteen strains required for the resolution IV experiments were constructed, so only 28 out of the 32 designed experiments could be performed. Although resolution IV designs with 32 experiments allow the estimation of main effects for up to 16 factors, we considered the impact of 11 factors on production. This granted some redundancy in the design and allowed the estimation of all main effects with 28 out of the 32 designed experiments. However, if more factors had been considered, the construction of all the required strains would have been necessary for the estimation of main effects. This limitation can be solved with the machine learning (ML) analysis of strain libraries generated using one-pot random transformation [345, 357]. Still, although ML can identify significant factors with an impact on production, quantifying the impact

of interactions between factors, critical for bio-process optimization, is not trivial [345]. Moreover, the randomization of strains and process conditions, even when mini-bioreactor systems are used [41, 42], increases the complexity of creating suitable datasets for ML.

Here we focused on the use of fractional factorial designs to find the optimal production conditions given a design space defined by the selected factors and their levels. To achieve this, a resolution IV design to identify factors with the strongest impact on production and the importance of interactions was employed. These factors were subsequently fixed, and a resolution V design was used to identify the significant interactions. Alternatively, the factors with the most important main effects could have been optimized beyond the original design space. While the pH variable was binary and the use of unbuffered media was recommended, response surface methods could have been employed to optimize the expression of PAL-C4H and the supplemented Phe concentration [79, 86]. In this case higher expression levels and Phe concentrations should have been tested to evaluate the existence of an optimum.

Summarizing, through a systematic evaluation of 11 factors, including genetic modifications and process parameters, we uncovered some of the interplay between genetic and environmental factors in pCA production. Moreover, we demonstrate the power of DoE to provide insights into factor effects and interactions for process optimization. By leveraging the strengths of experimental design and statistical analysis, we provide a framework to find key factors that impact bio-process performance that could be systematically applied to guide strain design as well as scale-up/down efforts.

Declaration of interest

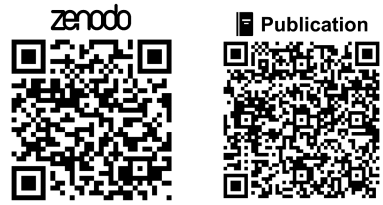
RvdH and JS are employed by dsm-firmenich, and VAPMdS has interests in LifeGlimmer GmbH.

Acknowledgment

This project was founded by the NWO (project number GSGT.2019.008) and the EU's Horizon 2020 research and innovation program (grant agreement 814408 (Shikifactory100)).

Data availability

Supplementary tables are available in Zenodo and the published version of the chapter.



CHAPTER 9

Machine learning-guided optimization of p-coumaric acid production in yeast

Sara Moreno Paz*, Rianne van der Hoek*, Elif Eliana, Priscilla Zwartjens, Silvia Gosiewska, Vitor A. P. Martins dos Santos, Joep Schmitz#, María Suárez Diez#

*Contributed equally, #Jointly supervised this work

This chapter is published in *ACS Synthetic Biology*
10.1021/acssynbio.4c00035

Abstract

Industrial biotechnology uses Design-Build-Test-Learn (DBTL) cycles to accelerate the development of microbial cell factories, required for the transition to a bio-based economy. To use them effectively, appropriate connections between each phase of the cycle are crucial. Using p-coumaric acid production in *Saccharomyces cerevisiae* as case study, we propose the use of one-pot library generation, random screening, targeted sequencing, and machine learning (ML) as links during DBTL cycles. We showed that the robustness and flexibility of ML models strongly enable pathway optimization, and propose feature importance and SHAP values as a guide to expand the design space of the original strain libraries. This approach allowed a 68% increased production of p-coumaric acid within two DBT(L) cycles leading to a 0.52 g/l titer and a 0.03 g/g yield on glucose.

Introduction

Climate change calls for an imminent transition to a bio-based economy less reliant on the petrochemical industry. Biotechnology contributes to solving this issue as metabolic engineering allows microbial production of a wide variety of compounds such as biofuels or bulk chemicals [24]. Yet, these solutions often require very long development times that limit their applications [28].

Design-Build-Test-Learn (DBTL) cycles offer a framework for systematic metabolic engineering. Pathways are designed during the Design phase, strains are constructed in the Build phase, and screened for production during the Test phase. In the Learning phase, a relationship between pathway design and production is established and used to inform new DBTL cycles [32]. Advances in synthetic biology and automation facilitate the engineering of microorganisms and increase the throughput of the Build and Test phases. However, predicting the effect of modifications in the Design phase that may lead to improvements is non-trivial [52, 98]. In fact, the acceleration of the Build and Test phases of the DBTL cycle might lead to a paradox where more data leads to more complexity but not necessarily better strain performance [50]. To avoid this, an efficient, meaningful link between the Design and Learn phases of the cycle is crucial.

Machine learning (ML) can identify patterns in the system of interest without the need for a detailed mechanistic understanding of the problem [358]. It has been used to aid strain development with applications ranging from gene annotation and pathway design to process scale-up [52]. When used for pathway optimization, common approaches start by creating libraries of strains with varying regulatory elements such as promoters or ribosome binding sites. These libraries include a defined solution space that can be explored by random or rational sampling [357, 359]. A subset of the library is then screened, and genotype and production data are used to train ML algorithms. The algorithms then suggest a new round of (improved) strains for construction, effectively linking the Learn and Design phases of sequential DBTL cycles [96, 97, 98, 357, 360]. Besides, ML algorithms are robust to missing data caused by unsuccessful construction of specific strains which facilitates effective and efficient implementation of DBTL cycles [96, 99].

p-Coumaric acid (pCA) is an aromatic amino acid-derived molecule produced from phenylalanine (Phe) or tyrosine (Tyr). It is naturally found in plants and serves as a starting material for commercially valuable products such as pharmaceuticals, flavors, fragrances, and cosmetics [340]. In *Saccharomyces cerevisiae* Phe and Tyr are synthetized via the prephenate pathway (Figure 9.1A) [46, 350]. This pathway starts with the condensation of erythrose-4-phosphate (E4P) and phosphoenolpyruvate (PEP) by 3-deoxy-7-phosphoheptulonate synthase (ARO3/4). Then, the pentafuntional protein ARO1 converts 3-deoxy-7-phosphoheptulonate (DAHP) to 5-enolpyruvylshikimate-3-phosphate (EPS3P), which is converted to chorismate (CHO) by ARO2, and to prephenate (PRP) by ARO7. Prephenate can then be converted to phenylalanine by prephenate dehydratase (PHEA) and ARO8/9, or to tyrosine by prephenate dehydrogenase (TYR) and ARO8/9. To continue the synthesis of pCA, expression of heterologous genes is needed: tyrosine ammonia lyase (TAL) for synthesis from Tyr; or phenylalanine ammonia lyase (PAL), cinnamate 4-hydroxylase (C4H) and its associated cytochrome P450 reductase (CPR) for synthesis from Phe [340, 341, 342, 343].

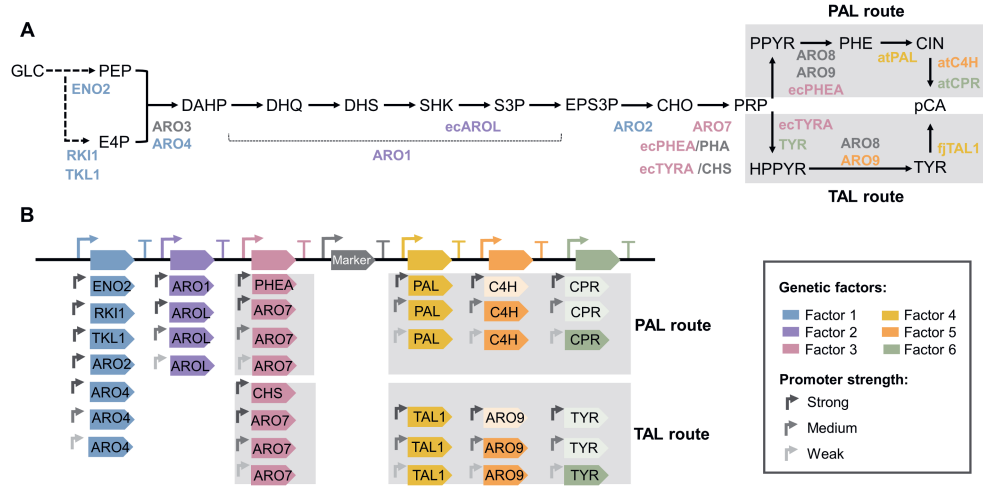


Figure 9.1: **A.** p-Coumaric acid (pCA) production pathway. Heterologous genes are preceded by a two-letter code indicating the organism of origin: ec, *Escherichia coli*; at, *Arabidopsis thaliana*; fj, *Flavobacterium johnsoniae*, see legend for color codes. **B.** Library structure. The library consists of gene clusters formed by a selection marker (Marker) and six factors with levels including different open reading frames and promoters. Lighter colors indicate factor levels included in the design but not obtained experimentally. GLC, glucose; PEP, phosphoenolpyruvate; E4P, erithrose-4-phosphate; DAHP, 3-deoxy-7-phosphoheptulonate; DHQ, 3-dehydroquinate; DHS, 3-dehydroshikimate; SHK, shikimate; S3P, shikimate-3-phosphate; EPS3P, 5-enolpyruvylshikimate-3-phosphate; CHO, chorismate; PRP, prephenate; PPYR, phenylpyruvate; PHE, phenylalanine, CIN, cinnamate; pCA, p-coumaric acid; HPPYR, 4-Hydroxyphenylpyruvate; TYR, tyrosine.

The prephenate pathway is highly regulated, Tyr exerts feedback inhibition on ARO3 and ARO7 and Phe on ARO4 [350]. This regulation together with the availability of precursors and appropriate expression of heterologous genes have been demonstrated to influence pCA production [340, 341, 342]. However, testing the effect of these factors individually might result in the exclusion of possible synergistic effects. Alternatively, combinatorial optimization of metabolic pathways can facilitate the search for optimal production albeit involving the construction and testing of an exponentially growing number of strains [357].

We used ML-guided DBT(L) cycles to improve pCA production in *S. cerevisiae*. We created combinatorial libraries based on the Tyr or Phe-derived pathways, that simultaneously altered expressed coding sequences and regulatory elements (promoters) (Figure 9.1B). We showed a better performance of the Phe-derived pathway, which was further optimized based on ML predictions. Following this strategy, we achieved a 68% improvement in production within two DBT(L) cycles and a final pCA titer of 0.52 g/l resulting in a 0.03 g/g yield of pCA on glucose. Although higher pCA yields up to 0.15 g/g have been previously obtained [340], this study is an example of the use of ML-guided DBTL cycles to systematize the generation of efficient strains.

Materials and methods

Organisms and media

S. cerevisiae strains were derived from CEN.PK113-7D and grown at 30°C in Yeast Extract Phytone Dextrose media with 2% D-glucose for transformations and pre-cultures (YEPhD, 2% Difco™ phytone peptone (Becton-Dickinson (BD), 1% Bacto™ Yeast extract (BD)). Verduyn Luttik minimal media with 2% glucose was used for production experiments [361]. When required, antibiotics were added to the media at appropriate concentrations: 200 µg/ml of nourseothricin (Jena Bioscience) and 200 µg/ml of geneticin (G418, Sigma-Aldrich). *Escherichia coli* DH10B (New England BioLabs) was used as cloning strain and grown at 37°C in 2XPeptone Yeast Extract media (2XPY, 1.6% tryptone peptone (BD), 1% Bacto yeast extract (BD) and 0.5% NaCl (Sigma Aldrich)). When required, antibiotics were added to the media at appropriate concentrations: 100 µg/ml of ampicillin (Sigma-Aldrich), and 50 µg/ml of neomycin (Sigma-Aldrich). Solid medium was prepared by the addition of Difco™ granulated agar (BD) to the medium to a final concentration of 2% (w/v).

Cassette construction

DNA templates for promoters and terminators [362] as well as open reading frames (ORF) were codon optimized according to Roubos et al. [363] and can be found in Sup. Table 1. Bricks were assembled into cassettes (promoter + ORF + terminator) via Golden Gate (using BsaI-HF v2.0 (NEB) and T4 DNA Ligase (Invitrogen)) into a backbone plasmid containing a 50 bp homologous connector sequence to facilitate *in vivo* recombination of the gene cluster as described in Verwaal et al. [346] (Sup. Figure 9.1). Golden Gate products were transformed into chemically competent *E. coli* DH10B. The Wizard® SV 96 Plasmid DNA purification system (Promega, Madison, WI, USA) was used for plasmid isolation. Cassettes were confirmed by PCR using Q5® High-Fidelity DNA polymerase (NEB) with primers from IDT and analyzed on a LabChip® GX Touch Nucleic Acid Analyzer (Perkin Elmer). Plasmids with correct fragment size were amplified by PCR using Q5® High-Fidelity DNA polymerase (NEB) and integration site flanks (50 bp homologous region) were attached to the first and the last cassette of the gene cluster (Sup. Figure 9.1). PCR products were purified using Promega Wizard® SV PCR Clean-Up kit and quantified using DropSense 96 (Trinean).

Strain construction

Strains were constructed as described in Verwaal et al. and Ciurkot et al. [346, 364]. In short, host strain SHK001 pre-expressing Cas9 (Sup. Table 3) was used to enable a targeted genomic integration of gene clusters Sup. Table 2. A linear guide RNA targeting a single locus (Sup. Table 6) was amplified from a gBlock (IDT) with 50 bp homology regions to the pRN1120 plasmid. PCRs were performed with Q5 ® High-Fidelity DNA polymerase (NEB). PCR products were confirmed on a 0.8% agarose gel and purified using Wizard® SV Gel and PCR Clean-Up kit (Promega). DNA fragments were quantified using Nanodrop. Plasmids and primers used are provided in Sup. Table 4 and 5.

Equimolar amounts (100-300 ng/kb) of the cassettes, linear gRNA (210 ng/kb), and linear backbone fragments (35 ng/kb) were transformed to the cells following the LiAc/ssDNA/PEG method [348]. Reagents for yeast transformation were obtained from Sigma-Aldrich (lithium acetate dihydrate (LiAc) and deoxyribonucleic acid sodium salt from salmon testes (ssDNA)) and Merck (polyethylene glycol 4000 (PEG)). *In vivo* recombination of the clusters is facilitated by connector sequences on the cassettes [346]. Transformants were plated on Qtray (NUNC) containing YEPhD agar medium and a selection agent. Colonies appeared on the plate after 3 days of incubation at 30°C. Single colonies were picked with Qpix 420 (Molecular Devices) into 96 well plates containing YEPhD agar medium and selection agent and regrown for 3 days at 30°C.

Whole genome sequencing

S. cerevisiae cells (OD 5-10) were pelleted and lysed in 200 µl 0.9% physiologic salt with 2 µl RNAse cocktail (Invitrogen) and 5mg/ml Zymolyase 100T (MP Biomedicals). The mixture was incubated at 37°C for 45 min. 200 µl 2X cell lysis solution (0.05M EDTA, 4%SDS) was added to the mixture and vortexed. 168 µl protein precipitation solution (10M NH4Ac) was added and proteins were precipitated by centrifugation for 10 min at 20K rcf at 4°C. The DNA in the supernatant was precipitated with an equal volume of isopropanol followed by centrifugation for 2 min at 16K rcf at room temperature. The DNA pellet was washed with 70% ethanol. The ethanol was discarded and the pellet was left to dry and then dissolved in MilliQ water. The isolated genomic DNA was quantified using Qubit (Thermofisher Scientific) and Nanodrop (Thermofisher Scientific), purified using Zymo Research gDNA Clean & Concentrator kit, and sequenced using the ligation sequencing kit (LSK-SQK109) with the native barcoding expansion (EXP-NBD114) from Oxford Nanopore Technologies according to manufacturer instructions on a GridION device (FLOW-MIN106 flow cell).

Promoter-terminator characterization

Combinations of promoter-terminators were characterized using GFP as a reporter gene. Precultures were prepared in 96-half-deep well plates (HDWP) containing 350 µl YEPhD + Pen/Strep (Invitrogen) and incubated at 30°C, 750 rpm, 80% humidity for 48 h. 10 µl of the grown pre-culture were re-inoculated to MTP-R48-B FlowerPlate (m2p-labs) containing 1 ml minimal medium + Pen/Strep (Invitrogen). The plate was incubated 48 h in Biolector® at 30°C, 800 rpm, 85% humidity. Biomass (em. 620nm/ex. 620nm) and fluorescence (em. 488nm/ex. 520nm), each with 3 filters (gain of 100, 50, and 20), were measured every 15 min. 40 µl of 2 days-old main culture were measured using fluorescence-activated cell sorting (BD, FACSAria Fusion) to detect single cells expressing GFP at a flow rate of 10,000 evt/s. The signal of fluorescent proteins was detected with a bandpass filter set at 530/30 nm for GFP. The data was recorded using BD FACSDiva 8.0.2 software to retrieve the geometric mean of the fluorescence distribution. Data was analyzed using FlowJo (v10.6.2).

p-Coumaric acid production experiments

Colonies were grown in 96 microtiter plates (MTP) Nunc flat bottom (ThermoFisher Scientific) containing YEPhD and appropriate selection agent for 48 h at 30°C, 750 rpm, 80% humidity. Cultures were re-inoculated in HDWP (ThermoFisher Scientific, AB-1277) containing 350 µl YEPhD and selection agent and grown for 48 h. The grown cultures were reinoculated to HDWP containing 350 µl minimal media and incubated for 2 days at 30°C, 750 rpm, 80% humidity. In all plates blank wells and wells containing a control strain (SHK0046, see Sup. Table 3) were included. For flow-NMR measurements, 250 µl broth were sampled to a 96-deep well plate (DWP) and mixed with 500 µl acetonitrile (Sigma Aldrich) by pipetting. The mixture was centrifuged at 4000 rpm for 10 min. 500 µl supernatant was transferred to a new DWP for analysis with flow-NMR. For LC/MS measurements 250 µl broth was sampled. 1 ml acetonitrile was added, the sample was mixed by pipetting and centrifuged. 250 µl supernatant was diluted with 375 µl milliQ and used for analysis with LC/MS.

p-Coumaric acid quantification with automated segmented-flow NMR analysis

The DWP plates were lyophilized to remove the non-deuterated solvents. 100 µl solution of 1 g/l internal standard 1,1-difluoro-1-trimethylsilanyl methylphosphoric acid (FSP, Bridge Organics) in MilliQ water was added into DWP prior to the lyophilization. To the lyophilized samples, 600 µl of D_2O (Cambridge Isotope Laboratories (DLM-4)) was added and homogenized. The samples were analyzed on a CTC PAL3 Dual-Head Robot RTC/RSI 160 cm robotic autosampler (CTC Analytics AG, Zwingen, Switzerland) fluidically coupled to a Bruker spectrometer Avance III HD 500 MHz UltraShield [365]. ^{1}H spectra were recorded with standard pulse program (zgcppr) with the following parameters: 16 scans, 2 dummy scans, 33k data points, 16.4 ppm spectral width, 1.2 s relaxation delay (d1), 8 µs 90° pulse, 2 s acquisition time, 15 Hz water suppression, and fixed receiver gain of 64. Spectra were processed and analyzed using Topspin 4.1.4 (Bruker). Spectral phasing was applied and spectra were aligned to 3-(trimethylsilyl)-1-propanesulfonic acid-d6 sodium salt (DSS-d6, Sigma-Aldrich) at 0 ppm. Auto baseline correction was applied on the full spectrum width. Additional third-order polynomial baseline correction for selected regions was applied if needed. The amount of pCA (doublet, 6.38 ppm, n=2H) was calculated relative to the signal of FSP. NMR production data per plate was normalized by the production of the SHK0046 control strain.

p-Coumaric acid quantification with LC-HR-MS spectrometry

Samples were analyzed on a Vanquish Horizon UHPLC system coupled to a Q Exactive Focus mass spectrometer (ThermoFisher). Chromatographic separation was achieved on an Acquity UPLC® BEH C18 column (100 x 2.1 mm, 1.7 µm, Waters), using gradient elution with 0.025% formic acid in LC-MS grade water (A), and 90% LC-MS grade acetonitrile (B). The gradient started with 1% B linearly increased to 50% B in 5 min, followed by an increase to 99% B in 0.1 min, kept at 99% for 1.9 min and then re-equilibrated with 1% B for 1.9 min. The flow rate was kept at 0.6 ml/min, using an injection volume of 2 µl, and the column temperature was set to 50°C. pCA was detected in

negative APCI mode and quantified using an external calibration line of a reference standard. Using this chromatographic system, the coumaric acid elutes at retention times 3.05 min with m/z 163.0403 (M-H), in good agreement (within 2 ppm) with the theoretical m/z value of 163.04007.

Machine learning-guided strain design

Originally, the PAL and TAL libraries each contained 3024 different designs. However, due to problems during cassette construction, the design space was reduced to 672 designs per library. We randomly screened 440 strains per library and classified them into four clusters based on NMR pCA production titers. Strains from every cluster were randomly selected for sequencing. Colonies were considered correct when they had targeted integration of the complete gene cluster (7 cassettes, one per factor, and the selection marker). For gene clusters present in more than one correct sequenced colony, average pCA production was considered. Two datasets were used: a *complete dataset* including data from producers and non-producers and a *producers dataset*. Colonies with pCA production below 0.05 a.u. were considered non-producers.

From the available regressor models in the scikit-learn library, the performance of multiple linear (MLR), support vector (SVR), random forest (RFR), and kernel ridge (KRR) regression models was evaluated. pCA production was modeled using the factor levels (genes or their expression strength), treated as categorical variables using one-hot encoding, as inputs (Table 9.1). Models were evaluated on their ability to predict pCA titers measured by NMR (model output). Each dataset was split into train (90% data) and test (10%) sets using stratification (*i.e.* maintaining the proportion of the different classes in both sets). For all models except MLR, hyper-parameters were selected based on leave-one-out cross-validation in the train set using the maximum error as score. Predictions of models with optimized hyper-parameters were compared to the test set using the coefficient of determination (R^2) as score. This process was repeated ten times and models were compared based on their average R^2 on the test sets. Additionally, the impact of the training data size on model performance was tested: after the train test split, percentages of the training data from 5 to 100% were used for training, and model performance was evaluated using the test set with R^2 as score. See Sup. Figure 9.2 for an overview of the model selection strategy.

For each dataset, models selected based on R^2 were trained following two different strategies: "*one-time*" and "*recurrent*" training. In the first strategy, all data from the dataset was used for training. In the second strategy, 90% of the data was used for training and this process was repeated 100 times. Trained models were used to predict pCA titers for all the designs in the design space. For each dataset and training strategy, top producers were ranked based on the frequency of each design being predicted as top 1, top 5, and top 10 by each model (Sup. Figure 9.2).

The impact of the different factors on pCA production was evaluated by permutation feature importance using the permutation_importance function of the scikit-learn library. In addition, SHapley Additive exPlanations (SHAP) values were calculated using the shap library [366].

All data and scripts used are available in Gitlab. Model selection, training, and feature importance were performed using Python (v3.8.8) and Scikit-learn (v1.1.3) [367].

Results

DBTL Cycle 1: Exploring the design space

Design: selection of factors and levels

Two independent libraries were designed depending on whether pCA was produced from Phe (PAL route) or Tyr (TAL route) (Figure 9.1A). Any design of the libraries is formed by a 7-genes cluster (6 factors and a selection marker) integrated in the genome of *S. cerevisiae* (Figure 9.1B, Sup. Figure 9.1). The combination of promoter, ORF, and terminator (cassette) in the gene cluster constitutes a factor that can take different levels depending on the chosen promoter and/or ORF. The size of the library is determined by the number of factors and levels so $library = \prod_{i=1}^{F} L_i$, where F is the number of factors and L_i is the number of levels of factor i. Both libraries shared factors 1 and 2 and differed in the other 4 factors (Figure 9.1B).

Factor 1 contained five ORFs: enolase (ENO1), ribose-5-phosphate isomerase (RKI1), transketolase (TKL1), ARO2, and a feedback-resistant ARO4 (ARO4^{K229L}) under the TDH3 promoter. Besides, ARO4^{K229L} could be downstream of two additional promoters (RPL8A, MYO4) as the expression of this gene has resulted in significantly increased pCA titers [340, 342]. ENO1, RKI1, and TKL1 were chosen considering that the availability of PEP and E4P can also affect production. ARO2 was included as an additional level to test the effect of other shikimate pathway genes (Table 9.1).

Levels for factor 2 were based on the assumption of ARO1 as rate-limiting step (Table 9.1). Rodriguez et al. observed increased pCA production when ARO1 or AROL from *E. coli*, which catalyzes the phosphorylation of shikimate, were over-expressed in yeast [342]. Therefore, 4 levels were chosen: expression of AROL under three different promoters (TEF1, RPL28, UREA3) and expression of ARO1 under a strong promoter (TEF1).

The focus of factor 3 was the expression of the feedback-resistant variant ARO7^{G141S} under three different promoters (PRE3, ACT1, PFY1), as overexpression of this gene improved pCA titers [340, 342]. Besides, the expression of PHEA and TYRA from *E. coli* with the PRE3 promoter are considered as additional levels for the PAL and TAL libraries respectively (Table 9.1). These bifunctional enzymes have a chorismate mutase activity and either prephenate dehydratase or dehydrogenase activities, specific for the formation of Phe or Tyr respectively [368, 369].

Factors 4, 5, and 6 of the PAL library each focused on one of the heterologous genes required for pCA production from Phe: PAL, C4H, and CPR under the control of three different promoters (ENO2, RPS9A, VMA6; KI_OLE1, CHOI, PXR1; and PGK1, RPS3, CCW12 respectively). In the TAL library, levels of factor 4 were formed by TAL under the control of three promoters (ENO2, RPS9A, and VMA6). In order to obtain a design space with the same size as the PAL library, factors 5 and 6 included the expression of ARO9 and TYR with the same promoters used for the PAL library (Table 9.1).

Considering the factors and levels used, the number of possible designs in each library was 3024 ($7 \cdot 4 \cdot 4 \cdot 3 \cdot 3 \cdot 3$).

Table 9.1: Summary of factors and their levels in the TAL and PAL libraries.

Factors	Levels (promoter + ORF)						
	1	2	3	4	5	6	7
1	TDH3-ENO2	TDH3-RKI	TDH3-TKL	TDH3-ARO2	TDH3-ARO4	RPL8A-ARO4	MYO4-ARO4
2	TEF1-ARO1	TEF1-AROL	RPL28-AROL	UREA3-ARO4			
3	PRE3-PHA PRE3-CHS	PRE3-ARO7	ACT1-ARO7	PFY1-ARO7			
4	ENO2-PAL ENO2-TAL	RPS9A-PAL RPS9A-TAL	VMA6-PAL VMA6-TAL				
5	KI_OLE1-C4H KI_OLE1-ARO9	CHO1-C4H CHO1-ARO9	PXR1-C4H PXR1-ARO9				
6	PGK1-CPR PGK1-TYR	RPS3-CPR RPS3-TYR	CCW12-CPR CCW12-TYR				

Build and Test: construction and screening of the combinatorial library

For each of the promoter-terminator pairs designed, cassettes formed by promoter-GFP-terminator were constructed and transformed into yeast. Positive colonies were found for all the constructs but the strong promoter-terminator pairs for factors 3 and 5 (PRE3-ADH1, OLE1-TDH3). Cells were grown in BioLector bioreactors and fluorescence was analyzed using FACS. For factor 1, the fluorescence of strong and medium promoters differed by an order of magnitude. For factors 2, 4 and 6, fluorescence values for the medium promoters were approximately half of those from strong promoters. Weak promoters showed fluorescence values 1 or 2 orders of magnitudes below the strong and medium promoters (Sup. Figure 9.3, Sup. Table 3).

Cassettes required for the *in vivo* assembly of the gene clusters were created by combining promoters, ORFs, terminators, and homology regions. All cassettes except those containing the strong promoter for factor 5 and the strong and medium promoters for factor 6 were obtained, which reduced the size of the PAL and TAL libraries from 3024 possible designs to 672 designs per library (Figure 9.1B).

S. cerevisiae cells expressing Cas9 were transformed with a mixture of the correct cassettes using one-pot transformation. Cells were plated in selective media and 440 strains per library were randomly selected for screening of pCA production. These stains were grown in 96 DWP for 48h, pCA was extracted and samples were measured using NMR (Figure 9.2). Colonies from the PAL route produced pCA ranging from 0 to 0.22 a.u. Colonies from the TAL route produced significantly less pCA, with only three colonies producing above the detection limit (0.05 a.u.) and a maximum production of 0.10 a.u.

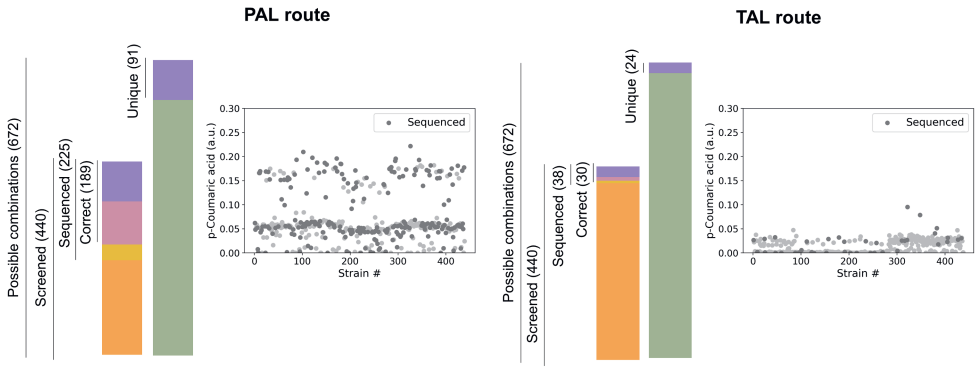


Figure 9.2: Screening before sequencing strategy. For each of the routes allowing pCA production, 672-member libraries were defined. For the PAL route, production of 440 randomly selected strains was measured and 225 strains were selected for sequencing; 189 correct strains containing 91 unique pathway designs were found. For the TAL route, 440 strains were screened from which 38 were sequenced; 30 of these strains were correct and 24 unique designs were found.

Considering the screening results, the production space was sampled including low, medium, and high producers to obtain high-quality data for ML and to analyze the efficiency of the library generation method. For the TAL route, 38 strains were sequenced from which 30 sequences were correct (*i.e.* contained the gene cluster with 7 genes) and 24 contained unique pathway designs (*i.e.* different integrated gene clusters) (Figure 9.2). Considering that 80% of the sequenced strains were correct, the observed low pCA production was likely caused by the lower efficiency of the TAL route and not incorrect library construction. These results agreed with previous reports that identified the PAL route as the most suitable pathway for pCA production [340]. Therefore, the optimization of pCA production was focused on the PAL route. Out of 672 possible designs in this library, 225 strains were selected for sequencing based on their pCA titers, ensuring that strains from different clusters were sequenced. We found 189 correct strains (84%) from which 91 (48%) were unique, validating the library construction approach (Figure 9.2). Out of the 91 unique designs, 58 designs were present in one strain and 33 had multiple replicates (Sup. Figure 9.4). Besides, for all factors at least a strain containing each of the levels was found.

Learn: model selection, training, and predictions

One of the challenges of applying ML to strain design is the training data requirements. While some reports suggest the homogeneous sampling of the complete solution space [52], others suggest the benefit of including mainly good producers [97]. Therefore, we divided our data into two datasets: the *complete* dataset that included data from producers and non-producers and the *producers* dataset. Stratification was used during training to ensure a constant proportion of poor, medium, good, and very good producers in the train and test sets. After hyper-parameter tuning, train sets were used to train four ML algorithms: MLR, SVR, KRR, and RFR. While MLR assumes

a linear relationship between the factors and the response, SVR and KRR can capture non-linear relationships, and random forest is an ensemble method that excels at handling complex interactions. The performance of the models with optimized hyper-parameters was evaluated on the test set. Models trained with the producers dataset showed better performance than those trained using the complete dataset (Table 9.2). MLR and KRR or all models were chosen as predictors for the *complete* and the *producer* datasets, respectively.

Table 9.2: Performance of ML methods (R^2) on test data. Models were trained with training data from the complete or producer datasets using stratification. MLR, multiple linear regression; SVR, support vector regression; KRR, kernel ridge regression; RFR random forest regression.

	Complete dataset (91 designs)	**Producers dataset** (63 designs)
MLR	0.70 ± 0.17	0.82 ± 0.15
SVR	0.71 ± 0.23	0.82 ± 0.19
KRR	0.72 ± 0.18	0.80 ± 0.11
RFR	0.72 ± 0.19	0.82 ± 0.16

Selected models were trained in each dataset using two different learning strategies: "*one-time training*" and "*recurrent training*" (Sup. Figure 9.2). The first strategy consisted of one-time training with all the available data and did not provide uncertainty in the predictions. The second strategy was based on recurrent learning on 90% of the available data which reduced the impact of possible outliers in the training data and allowed uncertainty quantification of predictions. Trained models were then used to predict the pCA titers of the 672 designs from the full design space. Considering that models selected for each dataset had similar performances (Table 9.2), designs were ranked based on the frequency in which each design was predicted to be in the top 1, top 5, or top 10 by each model. In this way, the construction of designs commonly predicted as top producers by different models was favored. Four rankings were obtained: the CO and CR rankings based on the Complete dataset and the One-time or Recurrent training strategies respectively, and the PO and PR rankings based on the Producers dataset (Sup. Figure 9.2).

Training strategies were evaluated based on their ranking of the best-measured producer strain (BMP), the five best-measured producers (5-BMPs), and all the measured non-producers (Sup. Figure 9.5). Best measured producers were expected to rank high while measured non-producers were expected to hold lower positions. Regardless of the training strategy, including non-producers during training did not change predictions of measured top producers, but improved predictions of measured non-producers, ensuring correct coverage of the complete design space by the ML predictions.

In order to improve pCA production, designs predicted to render the highest titers were evaluated (Sup. Figure 9.6). Notably, the BMP strain was predicted as part of the top 10 designs in all but the CO ranking. A comparison between the levels present in the training data and the top 10 designs predicted by all the learning strategies is depicted in Figure 9.3A and Sup. Figure 9.6.

Top predicted strains showed a preference for ARO4 under weak or strong promoters compared to the other ORFs. For factors 2 and 3, ARO1 or AROL under a strong promoter and PHA or ARO7 under its medium promoter were favored. Finally, the strongest promoters tested for PAL and C4H were enriched in the predicted top producers.

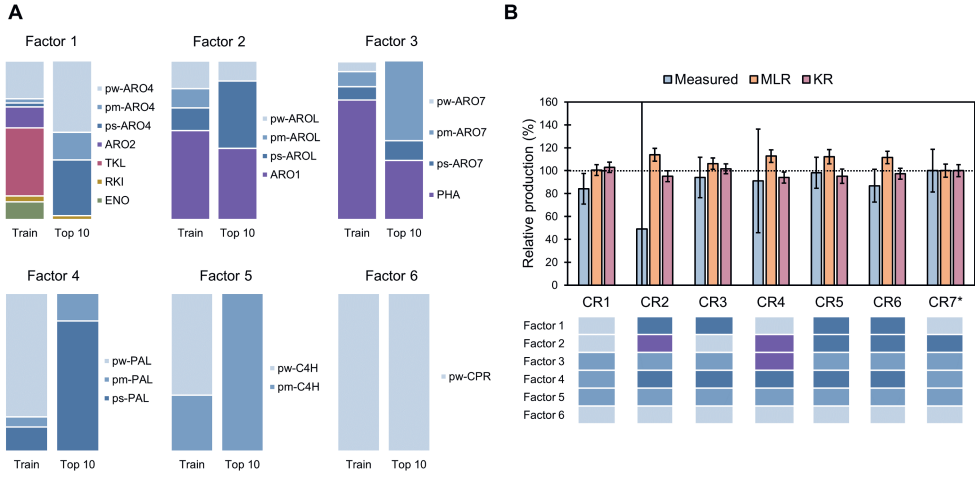


Figure 9.3: **A**. Comparison of factor levels on the training set and the top 10 predicted strains considering the CO, CR, PO, and PR rankings. Ps, pm, and pw indicate strong, medium, and weak promoters respectively. **B**. Experimental validation of the CR ranking predictions. Production relative to the BMP strain (same as CR7***** strain) is shown. The genotype of the strains follows the same color code presented in panel A. CO, complete dataset with one-time training; CR, complete dataset with recurrent training; PR, producers dataset with one-time training; PR producers dataset with recurrent training.

However, predicted pCA production improvements compared to the BMP strain were low ($6 \pm 8\%$, $2 \pm 5\%$, $3 \pm 6\%$, and $1 \pm 5\%$ depending on the learning strategy used, Sup. Figure 9.7). Therefore, we hypothesized that the initially screened library was a good representation of the whole design space and already achieved the highest possible production. Although we used 13.5% of the full library data for model training, ML algorithms are frequently trained with data representing circa 5% of the library design space. Therefore we tested the effect of reducing data availability on model performance (Sup. Figure 9.8, Table 9.3). When 40% of the available training data was used for training with stratification (equivalent to 5.4% of the library space), coefficients of determination of the test sets remained above 0.6 for all the models but MLR regardless of the dataset used, suggesting that the identification of 36 unique strains could have been sufficient for ML training.

Table 9.3: Performance of ML methods (R^2) on test sets when models are trained with training data size equal to 5.4% of the library. MLR, multiple linear regression; SVR, support vector regression; KRR, kernel ridge regression; RFR random forest regression.

	Complete dataset		**Producer dataset**	
	No stratification	**Stratification**	**No stratification**	**Stratification**
MLR	0.52 ± 0.24	0.56 ± 0.23	0.61 ± 0.25	0.65 ± 0.23
SVR	0.57 ± 0.25	0.65 ± 0.21	0.64 ± 0.27	0.73 ± 0.22
KRR	0.61 ± 0.22	0.67 ± 0.16	0.65 ± 0.21	0.64 ± 0.19
RFR	0.65 ± 0.22	0.70 ± 0.20	0.70 ± 0.24	0.75 ± 0.22

DBT Cycle 2: expansion of the original design space

ML analysis suggested that the optimal production possible considering the initial design space had already been found. In order to validate this prediction, top predicted designs by all the learning strategies were constructed. Figure 9.3B shows predicted and measured production of the top 7 designs in the CR ranking. As expected, the production of these strains did not significantly improve with respect to the BMP strain. Similar results were obtained with the top strains from the CO, PO, and PR rankings, with production remaining within the BMP mean $\pm$ 20% (Sup. Figure 9.9).

To improve pCA production, the original design space had to be expanded, and permutation feature importance and SHAP values were used to guide the new designs. Feature importance identifies the factors with the biggest influence on model performance by shuffling the levels of a factor and evaluating the decrease in model accuracy. Factor 5 (C4H expression) was identified as the most relevant factor, followed by factor 4 (PAL expression) (Figure 9.4A, Sup. Figure 9.11). Predicted top producers had C4H under the strongest promoter tested, and never chose the weaker promoter for PAL, suggesting higher expression of these genes could lead to higher production. This was confirmed by the SHAP values, a technique for explainable ML, that not only identifies significant factors but also determines how they affect the model output [366]. For all the training strategies used (except the MLR model with the producer dataset), the highest positive impact on model output was caused by expressing C4H and PAL under the strongest promoters. Similarly, the highest negative impact was caused by the expression of C4H and PAL under weaker promoters (Figure 9.4B, Sup. Figure 9.12). The importance of these genes was confirmed by substituting the promoters of PAL and/or C4H with weak promoters in the BMP strain and the best strain in the CR ranking (CR1). In both cases, strains with lower expression of PAL and/or C4H showed significantly reduced pCA tiers (Figure 9.4C). Besides, although unsuccessful cassette construction prevented testing the effect of different expression levels of CPR, changing the promoter of CPR in the BMP and CR1 strains did not significantly change pCA production (Sup. Figure 9.10).

To further increase the expression of the genes, strains with double copies of each gene were created using BMP and the best-constructed strain from the CR ranking (CR4) as hosts. Positive colonies containing double copies of ARO4 and ARO4-AROL-ARO7-PAL-C4H-CPR in the BMP host and PAL-C4H-CPR in the CR4 host could not be obtained. As expected, when extra copies of factors 1 (ARO4), 2 (AROL or ARO1), and 3 (ARO7 or PHEA) were integrated, production of pCA did not significantly change (Figure 9.4D). Production did not significantly increase either when double copies of PAL or C4H were integrated. Even though average production increased with a double copy of C4H, this change was not significant (Figure 9.4D). The integration of a double copy of the complete gene cluster was only achieved in one colony of the CR4 host, and its production was similar to strains with an extra copy of C4H. In both hosts, double copies of PAL and C4H resulted in significantly increased production (63% in BMP and 58% in CR4). Besides, significantly increased production was also found when double copies of PAL-CPR (36%) and C4H-CPR (60%) were expressed in CR4; and PAL-C4H-CPR were expressed in BMP (68%) (Figure 9.4D).

The observed increase in pCA production, only obtained when expanding the original design space, confirmed that the original space had been sufficiently sampled and validated feature importance and SHAP values as strategies to guide its expansion.

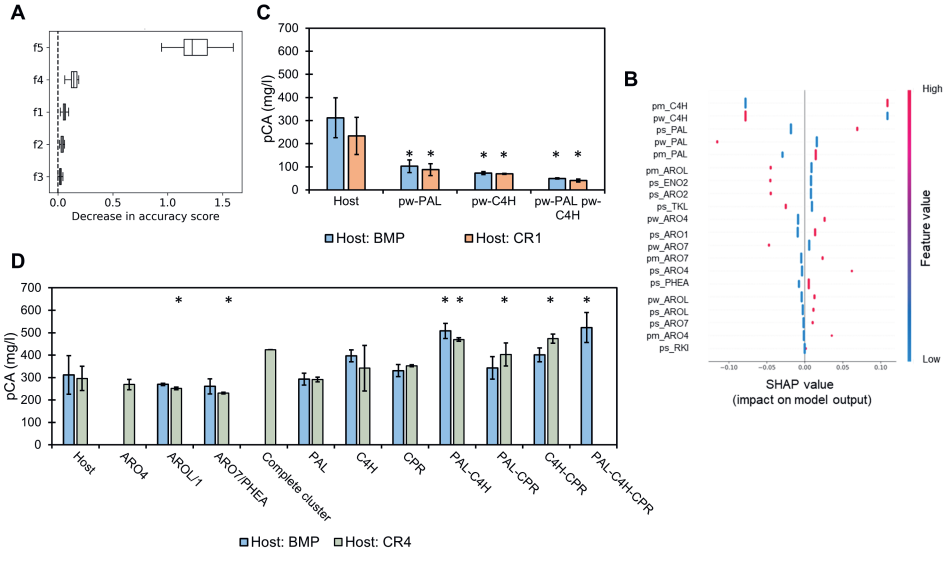


Figure 9.4: **A.** Representative example of feature importance results, where f1, f2, f3, f4, and f5 refer to factors 1 to 5. **B**. Representative example of SHAP values, where ps, pm, and pw refer to strong, medium, and weak promoters, respectively. **C.** Effect of substituting promoters of PAL and/or C4H by the weakest alternative (pw) in two different hosts: best-measured producer (BMP) and predicted top producer by the *complete recurrent* strategy (CR1). **D.** Effect of integration of double copies of genes in two different hosts: BMP and the constructed predicted top producer by the *complete recurrent* strategy (CR4). Significant differences compared to each host are indicated by *****. Colonies with double copies of ARO4 and the complete gene cluster were not obtained in the BMP host. Colonies with double copies of PAL-C4H-CPR were not obtained in the CR4 host and a single colony was obtained with the correct integration of the complete gene cluster.

Discussion

Accelerating the design of industrially relevant strains is crucial to transition to a bio-based economy. To exploit the full potential of microorganisms, combinatorial optimization of metabolic pathways is required. However, this involves the construction and testing of an exponentially growing number of strains which becomes unfeasible [89]. Alternatively, the solution space can be sampled following a rational or randomized approach. Statistical design of experiments reduces the number of strains to build and test while maximizing the information gained about the complete solution space. However, it requires the construction of specific strains and it is sensitive to experimental limitations: information is lost when a strain cannot be built [87]. As shown here, ML presents an alternative to learn from randomly generated libraries of strains which is robust to missing data. Besides, when ML is used, libraries can be flexibly designed to include factors with different number of levels based on prior knowledge. We used factor 1 to explore genes that could influence pCA production assigning 7 levels to this factor. Instead, we assigned 3 levels to factors 4, 5, and 6, aiming to fine-tune the expression of the required heterologous genes. We used 4 levels for factors 2 and 3 to simultaneously test the effect of homologous genes from different origins and tune the expression of one of them. The robustness and flexibility of the ML approach were also shown when some of the designed levels could not be implemented experimentally. Although the design spaces of the TAL and PAL libraries were reduced from 3024 members to 672, the relationship between the remaining levels could still be efficiently explored.

Another challenge to combinatorial pathway optimization is the need for the characterization of genetic parts that ensures that the solution space is sufficiently explored. This is especially important when the aim is to fine-tune the expression levels of pathway genes [96, 97, 98, 357]. In principle, the optimization of gene expression would benefit from the use of quantitative variables as factors (e.g. GFP fluorescence, protein levels) as they would allow the identification of an optimal expression level [333]. However, although effort is taken to appropriately characterize how regulatory elements affect gene expression, this is seldom achieved as *in vivo* expression depends on factors such as the downstream gene [370] or the gene order in an operon [32] and cannot be accurately predicted. Alternatively, regulatory elements can be treated as categorical variables reducing the impact of the characterization data [357]. This approach allowed us to include non-characterized promoters as members of the library and avoid a further decrease in the design space size. Besides, the use of categorical variables does not limit factor levels to differences in expression strength. As shown here, factors might include levels that represent differences in expression but also different ORFs, broadening the scope of ML-guided pathway optimization to the selection of genes from different origins or alternative over-expression targets.

A limitation to the use of ML is the requirement for sufficient and quality data for training [52]. We showed that including non-producers as part of the training set is not required to find top producing strains but improves predictions of poor producers which helps ensuring that the design space has been sufficiently sampled. This is especially important when the top producer is already present in the training data. Although we trained ML models with data representing 13.5% of

the library, we showed that when the amount of data used for training decreased, stratification during training improved the mean R^2 and reduced its standard deviation (Figure S12, Table 9.3). Stratification allowed the classification of samples based on production. Therefore, a sufficient number of samples from each category should be present in the training data. As shown here, this can be achieved using a screening before sequencing approach, which allowed an efficient exploration of the design space and reduced the chance of sequencing duplicate designs.

ML algorithms cannot predict the performance of strains with factor levels different from those used during training [358]. Still, they can be used to determine whether the best producer from the library is present in the training data and justify the expansion of the original design space. When this is required, feature importance and SHAP values can be used to guide this expansion and point at the most relevant factors which, in this case, led to a 68% improvement in pCA production. Notably, while feature importance only points as the significant factors, SHAP values provide additional information regarding how the factor's levels influence the model output [366].

The highest titers of pCA measured in this study were 0.51 ± 0.03 and 0.52 ± 0.06 g/l obtained using the BMP strain with additional copies of PAL-C4H or PAL-C4H-CPR. These strains were cultivated in 96DWP with minimal media and 20 g/l of glucose resulting in 0.03 g/g pCA yield. However, higher titers and yields of pCA have been reported. Rodriguez et al. obtained 1.96 g/l of pCA (0.04 g/g) by expressing AROL, feedback-resistant variants of ARO4 and ARO7, eliminating competing metabolic pathways and using synthetic fed-batch media [342]. Production was further improved by Liu et al. by combining the TAL and PAL pathways and including a phosphoketolase pathway to increase E4P availability. This strain produced 3.1 g/l in shake flasks and up to 12.5 g/l in bioreactors operated as fed-batch with a maximum yield of 0.15 g/g [340]. Considering these results, the production of our developed strains could be further improved in the next cycles that focus on gene deletions and media and bio-process optimization. This optimization would benefit from an improved experimental throughput achievable, for instance, using barcode sequences to mitigate sequencing costs [371] or a pCA biosensor for titer estimation [372]. This throughput, in turn, could allow the simultaneous testing of gene deletions, process conditions, and gene overexpression using multiple gene copies that could lead to further increased production. However, a trade-off between the build and test capacity and efficiency and the complexity of the learning step must be established by ensuring that a minimum percentage of the library space (e.g. 5%) can be used for model training. When this throughput is not achievable, sequential DBTL cycles, as those presented here, are useful to identify the relevance of the tested factors and levels and decide whether they are maintained or replaced in subsequent optimization cycles.

This study is an example of how ML-guided DBTL cycles can accelerate the generation of efficient strains. This approach is robust to experimental limitations and its flexibility regarding design, which can be expanded beyond the traditional tuning of gene expression. We propose a screening before sequencing approach to allow for stratification during training, especially important for small datasets. Furthermore, we showed how feature importance and SHAP values can be used to expand the original design space and further improve strain performance.

Declaration of interest

RvdH, PZ, SG and JS are employed by dsm-firmenich, VAPMdS has interests in LifeGlimmer GmbH.

Acknowledgment

This project was founded by the Netherlands Organization for Scientific Research (NWO; project number GSGT.2019.008) and the European Union's Horizon 2020 research and innovation program under grant agreement 814408 (Shikifactory100). Additionally we would like to thank Moniek Jonkers for her help with the laboratory automation workflows and Lieke Meijvogel, Sharina Chander, Judith Vis, Sylvana Suisse, Wibo B. van Scheppingen and Leon Coulier for execution and support with the analytical workflows.

Data availability

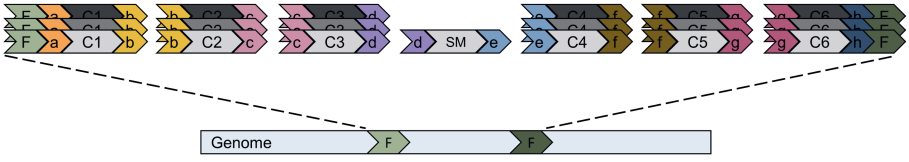
Scripts, data, and supplementary tables are available at Gitlab, Zenodo, and the published version of the chapter.

Supplementary Figures

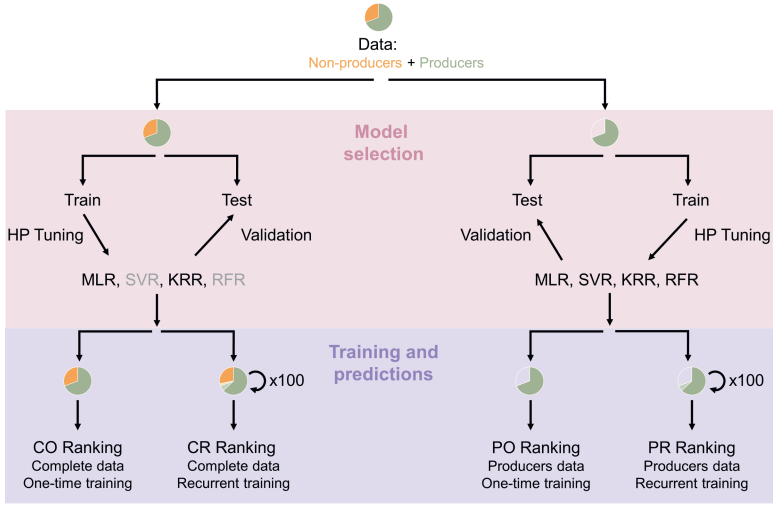
A

Factors	Possible promoters	Levels (promoter + ORF)							Terminator	# Cassettes (Prom + ORF + Ter)
		1	2	3	4	5	6	7		
1	ps = TDH3 pm = RPL8A pw = MYO4	ps-ENO2	ps-RKI	ps-TKL	ps-ARO2	ps-ARO4	pm-ARO4	pw-ARO4	TEF2	7
2	ps = TEF1 pm = RPL28 pw = UREA3	ps-ARO1	ps-AROL	pm-AROL	pw-AROL				PGK1	4
3	ps = PRE3 pm = ACT1 pw = PFY1	ps-PHA or ps-CHS	ps-ARO7	pm-ARO7	pw-ARO7				ADH1	4
4	ps = ENO2 pm = RPS9A pw = VMA6	ps-PAL or ps-TAL	pm-PAL or pm-TAL	pw-PAL or pw-TAL					TDH1	3
5	ps = KI_OLE1 pm = CHO1 pw = PXR1	ps-C4H or ps-ARO9	pm-C4H or pm-ARO9	pw-C4H or pw-ARO9					TDH3	3
6	ps = PGK1 pm = RPS3 pw = CCW12	ps-CPR or ps-TYR	pm-CPR or pm-TYR	Pw-CPR or pw-TYR					GPM1	3

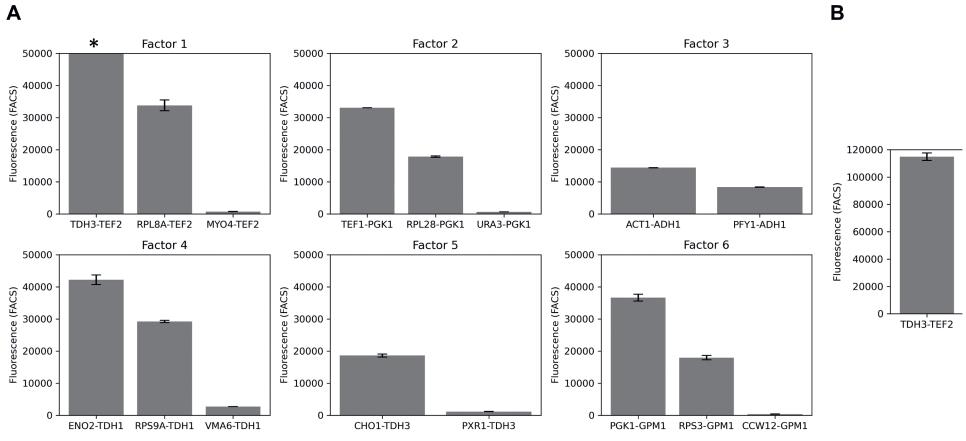
B



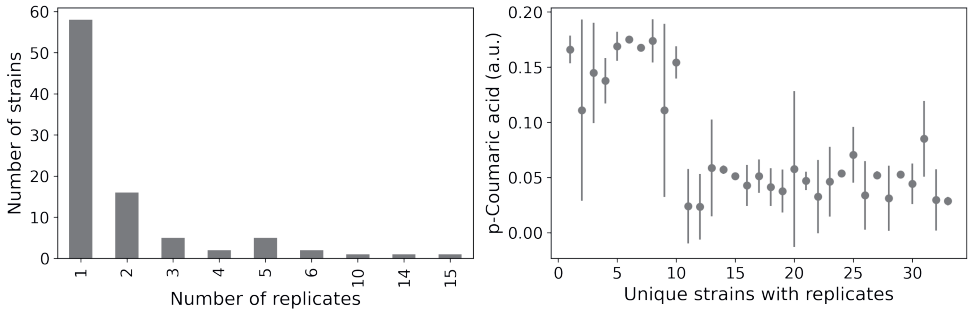
Sup. Figure 9.1: **A.** Cassettes used for library transformation. A cassette is a combination of promoter, open reading frame (ORF), and terminator. Cassettes containing promoter-ORF combinations shown in orange could not be obtained. **B**. Schematic representation of the integration of a gene cluster. Connector sequences a to h represent homology regions for *in vivo* recombination of cassettes (C1 to C6) and the selection marker cassette (SM); flank sequences homologous to the genome integration site are shown as F.



Sup. Figure 9.2: Model selection and training strategies. Genotype and production data were divided into two datasets: the *complete* and *producer* datasets which differ in the inclusion of data from non-producers. Each dataset was used for hyper-parameter (HP) tuning of four ML models: multiple linear regressor (MLR), support vector regressor (SVR), kernel ridge regressor (KRR), and random forest regressor (RFR). The accuracy of models with optimal HP was evaluated on the test sets. For each dataset, two learning strategies were applied: *one-time training*, where all the data was used for training, and *recurrent training*, where 90% of the training data was iteratively used for training.



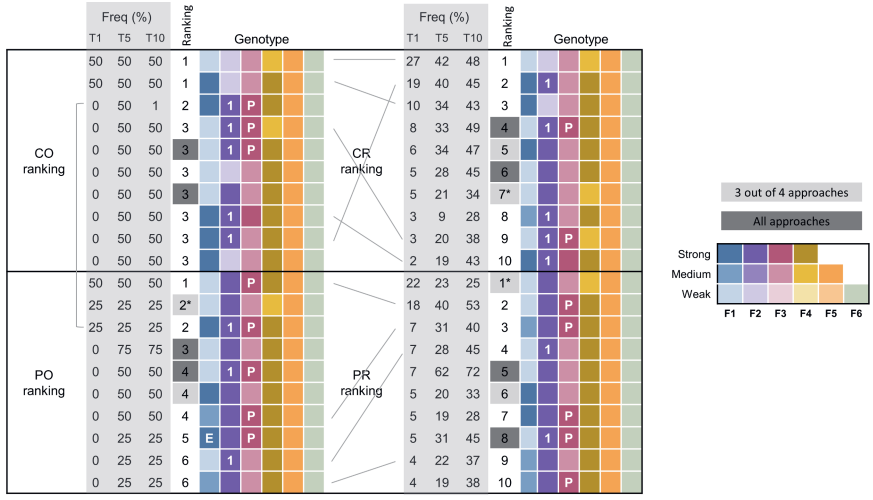
Sup. Figure 9.3: **A.** Promoter-terminator characterization by GFP fluorescence measured using fluorescence-activated cell sorting (FACS). **B.** Zoom-out for the fluorescence values.



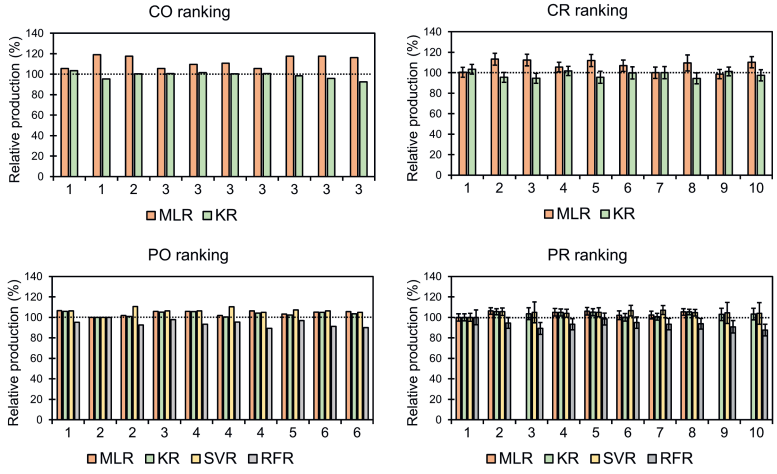
Sup. Figure 9.4: Characterization of strains with the correct sequences from the PAL library. Frequency of strains with the same designs (left) and average pCA production of designs with replicates (right).

	Best measured producer	Best 5 measured producers	Non-producers
CO Ranking	Predicted to be in: - Top 12% (MLR) - Top 0.9% (KRR)	Predicted to be in: - Top 19% (MLR) - Top 20% (KRR)	Predicted to be in: - Bottom 46% (MLR) - Bottom 40% (KRR)
CR Ranking	Predicted to be in: - Top 8% (MLR) - Top 2% (KRR)	Predicted to be in: - Top 18% (MLR) - Top 19% (KRR)	Predicted to be in: - Bottom 44% (MLR) - Bottom 40% (KRR)
PO Ranking	Predicted to be in: - Top 5% (MLR) - Top 15% (SVR) - Top 4% (KRR) - Top 0.1% (RFR)	Predicted to be in: - Top 23% (MLR) - Top 21% (SVR) - Top 22% (KRR) - Top 15% (RFR)	Predicted to be in: - Bottom 52% (MLR) - Bottom 60% (SVR) - Bottom 59% (KRR) - Bottom 60% (RFR)
PR Ranking	Predicted to be in: - Top 6% (MLR) - Top 10% (SVR) - Top 4% (KRR) - Top 3% (RFR)	Predicted to be in: - Top 23% (MLR) - Top 20% (SVR) - Top 23% (KRR) - Top 15% (RFR)	Predicted to be in: - Bottom 50% (MLR) - Bottom 52% (SVR) - Bottom 58% (KRR) - Bottom 61% (RFR)

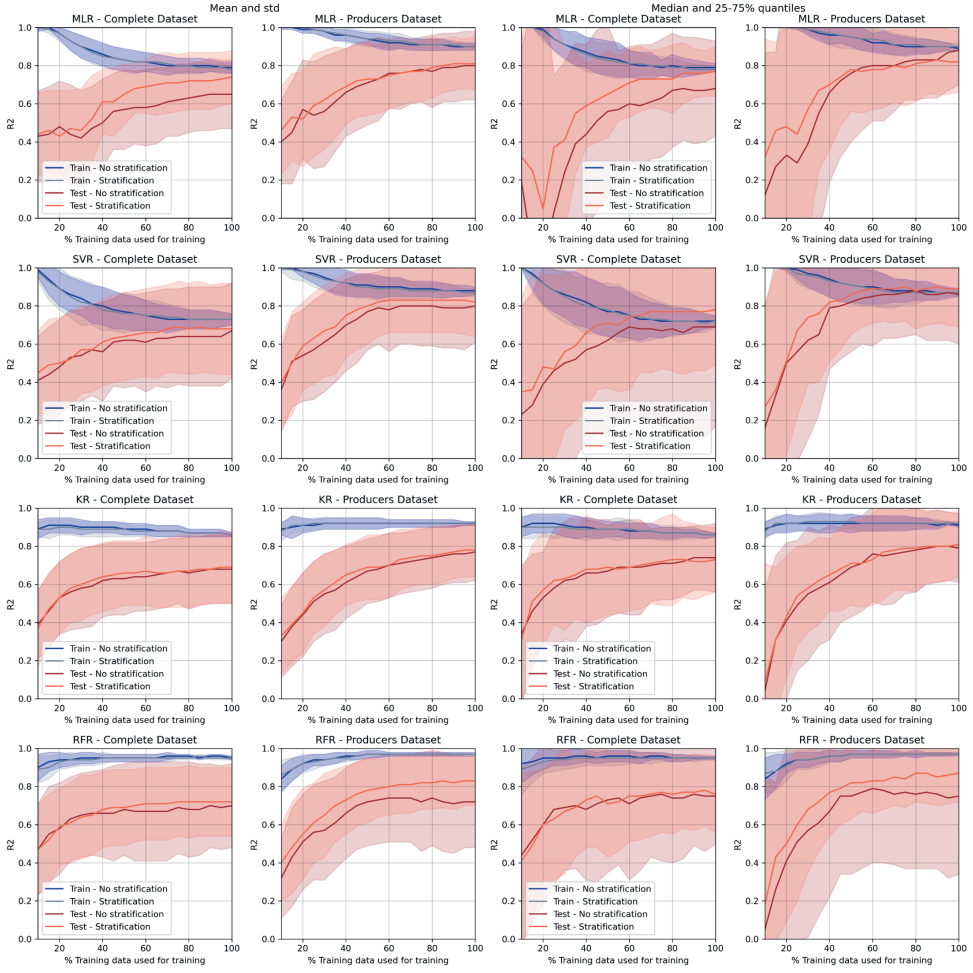
Sup. Figure 9.5: Ranking of top measured producers, top 5 measured producers, and non-producers based on four different training strategies. Results are given per model used in each strategy: MLR, multiple linear regressor; KRR, kernel ridge regressor; SVR, support vector regressor; RFR, random forest regressor. *Ranking of non-produces excludes design 560 which is predicted to produce by all models independently of the training strategy. The BMP was ranked in the top 0.1% to 15% depending on the model used and regardless of the training strategy. Similarly, the 5-BMPs were always predicted to be, at least, in the top 22% of the library. Measured non-producers were ranked in the bottom 46% or 60% of the library depending on the dataset used for training (*complete* or *producers* respectively). CO, complete dataset with one-time training; CR, complete dataset with recurrent training; PR, producers dataset with one-time training; PR producers dataset with recurrent training.



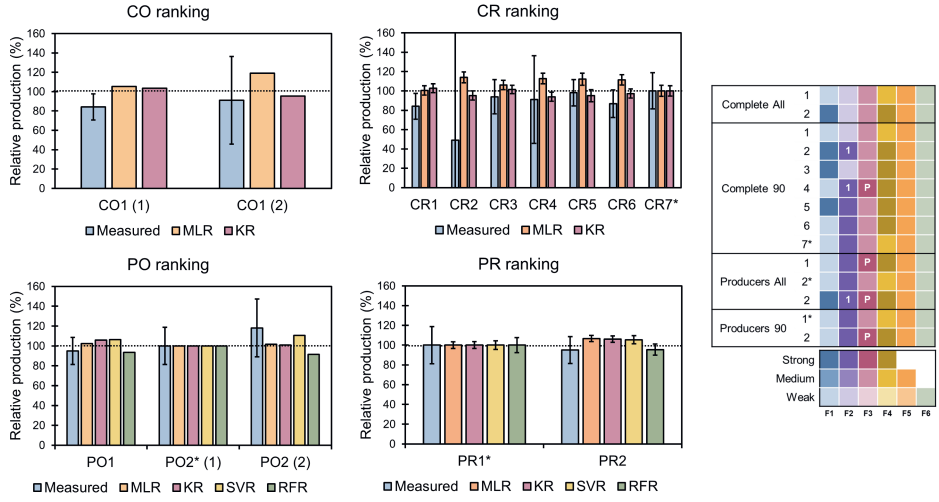
Sup. Figure 9.6: Summary of top 10 predicted producers by each learning strategy: CO, complete dataset with one-time training; CR, complete dataset with recurrent training; PO, producers dataset with one-time training; PR producers dataset with recurrent training. Designs are ranked based on the frequency (Freq.) they are chosen as top 1 (T1), top 5 (T5), or top 10 (T10) by the different models. Factor 1 (F1) refers to ARO4 except an E is shown (ENO1), factor 2 (F2) refers to AROL except a 1 is shown (ARO1), factor 3 (F3) refers to ARO7 unless a P is shown (PHEA), factor 4 (F4) refers to PAL, factor 5 (F5) refers to C4H and factor 6 (F6) refers to CPR. Predicted designs shared by different learning strategies are linked by lines or highlighted in grey. * Indicates designs equal to the best-measured producer strain (BMP strain).



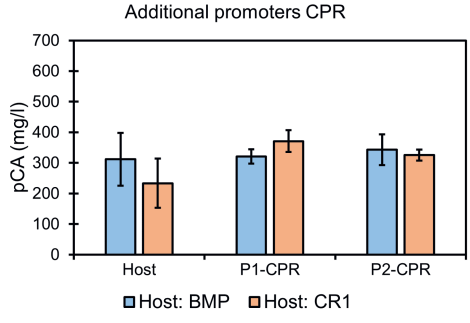
Sup. Figure 9.7: Predicted pCA production by the top 10 ranked strains found using four different learning strategies. Production relative to the predicted production of the top measured producer is shown. CO, complete dataset with one-time training; CR, complete dataset with recurrent training; PR, producers dataset with one-time training; PR producers dataset with recurrent training.



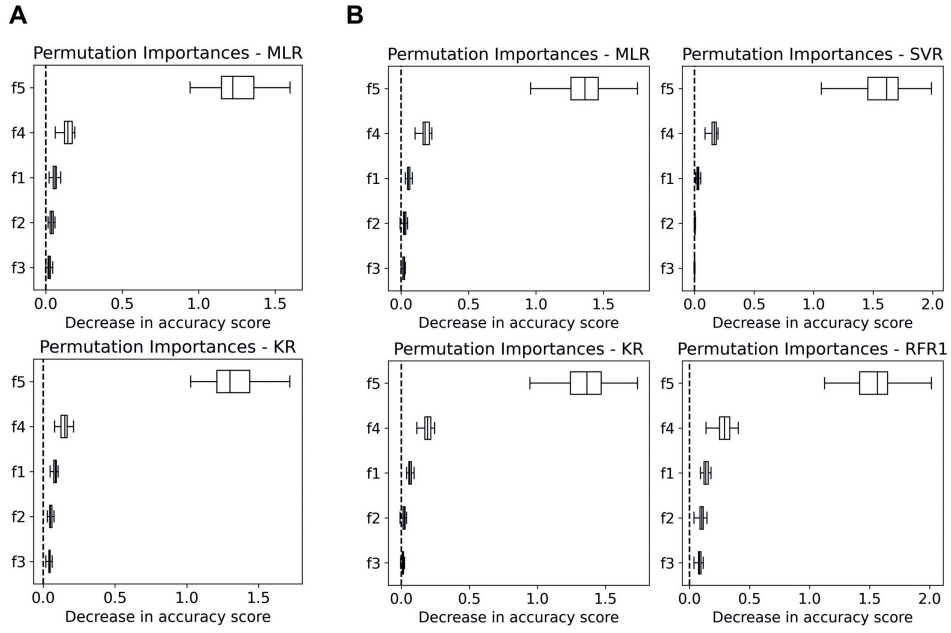
Sup. Figure 9.8: Effect of training data size on accuracy of predictions of different ML algorithms (MLR, multiple linear regression; SVR, support vector regression; KRR, kernel ridge regression; RFR, random forest regression) with the complete or producers datasets. Negative R^2 values obtained for some test-train splits were omitted for the calculation of mean and std (these values are obtained when the average of the training data is a better estimator than the trained model).



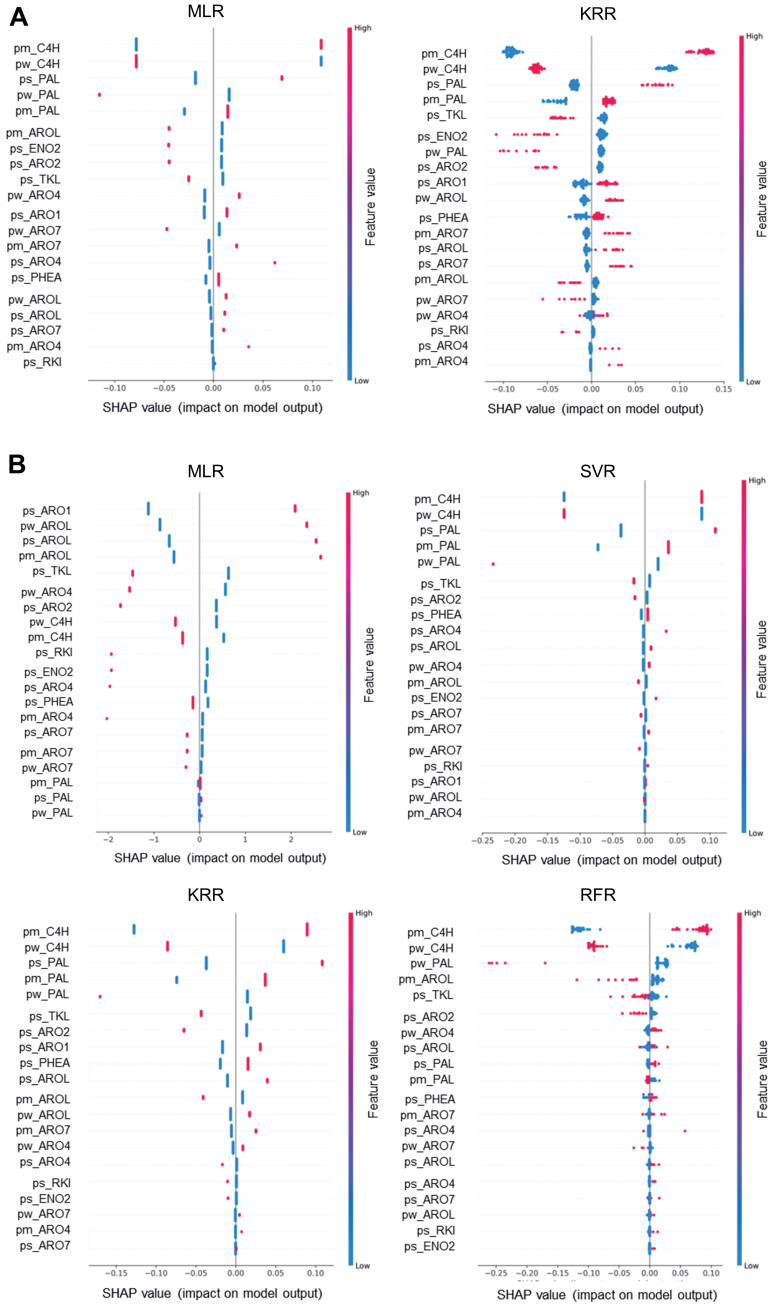
Sup. Figure 9.9: Validation of ML predictions. Comparison of measured and predicted production of the predicted top producers relative to the production of the best-measured producer (BMP). Genotypes of the plotted strains are shown in the right panel where * indicates strains equal to BMP. Factor 1 (F1) refers to ARO4, factor 2 (F2) refers to AROL except when 1 is shown (ARO1), factor 3 (F3) refers to ARO7 except when P is shown (PHEA), factor 4 (F4) refers to PAL; factor 5 (F5) to C4H, and factor 6 (F6) to CPR. Promoter strengths are represented by color intensity. Strain names are defined based on the ranking they belong to and their position in the ranking, when two strains share the same position they are followed by (1) and (2). CO, complete dataset with one-time training; CR, complete dataset with recurrent training; PR, producers dataset with one-time training; PR producers dataset with recurrent training.



Sup. Figure 9.10: Effect of substituting the CPR promoter in two different hosts: the best-measured strain (BMP) and the top producer in the CR ranking.



Sup. Figure 9.11: Permutation feature importance results obtained using the *complete* (**A**) or the *producers* (**B**) datasets. f1, f2, f3, f4, and f5 refer to factors 1 to 5



Sup. Figure 9.12: SHAP values obtained using the *complete* (**A**) or the *producers* (**B**) datasets. ps, pm, and pw refer to strong, medium, and weak promoters, respectively.

CHAPTER 10

General Discussion

Sara Moreno Paz

Preface

The objective of this thesis was to deploy various modeling methods for guiding and accelerating the design of cell factories and bioprocesses. The foundational elements of this work encompass four modeling approaches: kinetic modeling (**Chapter 2**), constraint-based modeling (**Chapters 3, 4, 5**), design of experiments incorporating linear regression and statistical analysis (DoE) (**Chapters 7, 8**), and machine learning (ML) (**Chapter 9**). While kinetic and constraint-based modeling rely on a mechanistic understanding of the system under study, DoE and ML are data-driven strategies for optimizing the analyzed system. Furthermore, omic analysis is a valuable tool for obtaining a comprehensive overview of the investigated system, especially when minimal information is available (**Chapter 6**). In this section, I will evaluate the employed modeling strategies and reflect on the strain and bioprocess design stages where their application is most pertinent. Then, I will ponder about the interplay between optimization and understanding in biotechnological research. Subsequently, I will examine the emergence of biofoundries and the associated challenges and opportunities of harnessing robotic platforms, computer-aided design, and human knowledge. I will conclude by discussing to what extent automation has the potential to fulfill biotechnology's promises and highlight the imperative need for critical sustainability assessments of biotechnological processes.

Modeling along DBTL cycles

Many factors affect the performance of cell factories and, in general, bioprocesses, rendering them difficult to understand and steer [14, 25]. This becomes even more challenging when considering that most of these factors are not independent and should, therefore be simultaneously evaluated. A pathway optimized in one organism might not be optimal when a different organism is employed or even when the metabolism of the original organism, or the expression of one of the pathway's genes is modified [340]. Moreover, the behavior of the microorganism also depends on the environmental conditions, and a process that performs well at lab-scale will most often suffer limitations at larger scales [25].

Factors affecting bioprocesses can be optimized using one factor at a time (OFAT) experimentation. When this approach is followed factors are optimized individually holding all other experimental variables constant [332]. However, as the interplay between different factors cannot be ignored, OFAT experimentation often leads to suboptimal strains and processes. For example, if the expression of a gene is limiting, the optimal expression of the other genes will only be found as long as the expression of the limiting gene is high. Combinatorial optimization, based on the simultaneous evaluation of multiple factors, captures these interactions, can better guide the optimization process, and must be prioritized over OFAT experimentation. This ensures the correct identification of the individual effect of a variable on the response as well as the effect of multiple factor interactions [73, 332]. However, evaluating all the factors that affect the performance of strains and bioprocesses simultaneously requires an experimental throughput currently unavailable. Hence, although combinatorial experimentation is essential, iterative experimentation is needed.

Design-Build-Test-Learn (DBTL) cycles are a good approach to tackle the iterative experimentation process. Although experimentation has to be sequential, DBTL rounds centered on the optimization of a pathway, the cell's metabolism, or the bioprocess can be alternated and combined without any specific order [339]. For instance, strains can be first optimized at the pathway and metabolic levels followed by process optimization. If at that point, new limitations are found, they could be solved by re-engineering the strain.

At any of these stages, the models explored during this thesis can help the researcher with the decision-making process. For example, if a kinetic model of the studied pathway is available (**Chapter 2**) it could be used to select which pathway genes are important and which expression levels are relevant. At the same time, methods based on genome scale metabolic modeling such as CFSA (**Chapter 3**) could provide a list of relevant metabolic engineering targets. Then, random combinations of strains containing (or not) these targets or expressing them at different strengths could be created and analyzed using machine learning (**Chapter 9**). In this way, the most important enzymes could be determined and further explored in combination with environmental factors, at a lower throughput, using statistical experimental design (**Chapter 8**). Strain engineering and bioprocess design are highly dynamic practices and, in this section, I focus on modeling approaches that prove valuable at various stages of this process (Figure 10.1).

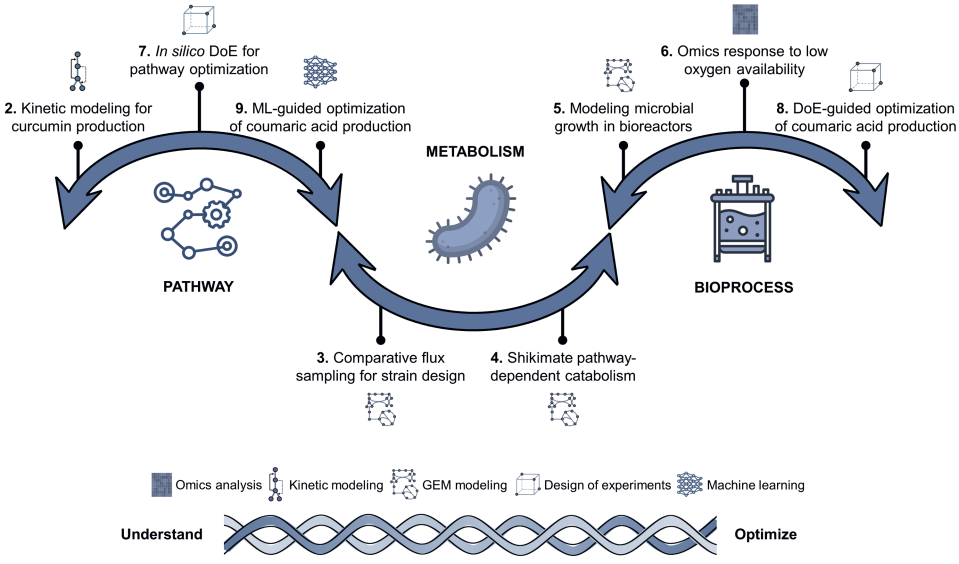


Figure 10.1: Thesis overview. Chapters are organized based on their focus on different stages during strain and bioprocess development. Modeling approaches used in the chapters are ordered on a scale considering their relevance to understanding and/or optimizing the studied system. DoE, design of experiments; ML, machine learning; GEM, genome-scale metabolic models. Pathway, metabolism, and bioreactor icons were obtained from freepik.com.

A focus on pathways

Chapters 2, 7, and 9 aimed at optimization at the pathway level: curcumin production in *P. putida* with the aid of a kinetic model, *in silico* curcumin production based on various DoE designs, and p-coumaric acid production in *S. cerevisiae* employing machine learning (Figure 10.1).

In **Chapter 2**, *P. putida* strains expressing the curcumin pathway were constructed and the acquired production data was used to create ensemble kinetic models of the pathway. The curcumin pathway involves several promiscuous enzymes capable of catalyzing competing reactions, which complicates the determination of optimal gene expression levels for the pathway genes. Strains were constructed to incrementally expand the range of possible substrate utilization by introducing new genes, and experiments with different strains and substrates were conducted. This approach aimed to facilitate model parameterization and deepen the mechanistic understanding of the pathway. However, the acquired data was not sufficient to obtain accurate parameter estimates and, instead, ensemble models were used. Even when not fully parameterized, models can serve as hypothesis generators [118], and the use of ensemble models allows the study of system properties [211, 373]. In **Chapter 2** ensemble models were enough to understand the pathway dynamics and obtain a 4.1-fold improvement in curcumin yield from tyrosine.

The approach followed in **Chapter 2** was fundamentally different from that proposed in **Chapters 7 and 9**. **Chapter 7** delved into the use of fractional factorial designs for the optimization of gene expression in a production pathway. Based on a predetermined design, strains were constructed and tested *in silico* and specific information about the system, including main effects and interactions between factors, was obtained. Alternatively, in **Chapter 9** a random library of *S. cerevisiae* strains was constructed and analyzed using ML. In both cases, mechanistic understanding of the studied system was not required, and the only prerequisite was the selection of factors with a potential effect on production and tunable levels. Notably, DoE and ML are complementary techniques and ML models can be trained with data generated by DoE designs [374]. Similarly, random data can be used to train multiple linear regression models (**Chapter 9**, [95]). However, as demonstrated in **Chapter 7**, the use of experimental data generated by DoE designs resulted in better linear models than those trained with random data. The decision to construct rational or random libraries depends on the ease of library construction, the desired information gain, the required flexibility for factors types and levels, and the experimental capabilities. For instance, while *in vivo* recombination in yeast is highly efficient [357], construction of random libraries in bacteria such as *E. coli* requires a more tedious process that starts with *in vitro* library construction and is followed by transformation [375]. Moreover, random libraries usually contain a larger number of strains which also requires high-throughput strain characterization facilities [98, 357].

In addition to the fractional DoE approaches explored in **Chapter 7**, other DoE designs that allow different numbers of levels per factor are available [32, 76]. Yet, ML provides higher flexibility in testing various factors and their levels, which allowed the exploration of different levels of gene expression and the expression of different genes in **Chapter 9**. Moreover, in this chapter, the number of levels per factor was selected based on the expected effect on production. However, this flexibility comes at the expense of increased difficulty in model interpretability. Although feature importance and SHAP values aided in the interpretation of these models, elucidating interactions among factors from ML models remains challenging [100, 366]. When identifying these interactions is a priority, the DoE strategies discussed in **Chapter 7** are more suitable, as DoE provides known information gains after experimentation. Alternatively, ML algorithms designed to detect interactions could be employed [376]. When possible, the use of these "transparent" models should be prioritized over the use of "opaque" models and post-hoc explainability techniques such as local approximations, model simplification, or feature relevance explanations [102].

Last, DoE and ML can also be employed to facilitate the parameter estimation process in kinetic models. For instance, optimal experimental design approaches select experiments that reduce the variance of estimated parameters in mechanistic models [78, 377]. Following the curcumin pathway example, Van Rosmalen et al. found that fermentations with multiple substrates could improve the estimation of kinetic parameters compared to single substrate experiments [77]. ML can, in turn, be used to accelerate the parameterization process of kinetic models [378] or to solve more complex tasks such as inferring the structure of a biochemical network from data [379].

A focus on metabolism

While Genome-scale Metabolic Models (GEMs) are not suitable to model pathways predominantly determined by enzyme kinetics, as the curcumin pathway explored in **Chapter 2**, they serve as the primary tool for guiding metabolic engineering at the genome scale. Constraint-based modeling has been used to successfully design strains to overproduce metabolites such as succinic acid, lycopene, or L-valine [24]. **Chapters 3 and 4** leveraged this tool to design cell factories employing two competing strategies: growth-uncoupled and growth-coupled production (Figure 10.1). Growth-uncoupled production, explored in **Chapter 3**, separates growth and production phases to alleviate metabolic stress, improving growth and production rates [160]. This separation can be achieved through operational changes, like nutrient limitations [274], or dynamic metabolic regulation based on metabolite sensing and gene circuits [159]. On the other hand, growth-coupled production, as investigated in **Chapter 4**, is based on the necessity to produce the target metabolite for growth, and is often combined with Adaptive Laboratory Evolution (ALE) [380, 381]. While growth-uncoupled production directs substrate use towards metabolite production, it can lead to phenotype loss due to the accumulation of mutations. Instead, growth-coupled strategies ensure sustained production over longer time frames at the expense of biomass generation as by-product.

In **Chapter 3** Comparative Flux Sampling Analysis (CFSA) exemplified how GEMs can streamline strain design. Despite various tools available for this purpose [152, 153, 154, 155], we focused on developing a method to design growth-uncoupled production strategies based on flux sampling, ensuring the complete exploration of the GEM's solution space. Although CFSA ranks reaction targets using multiple criteria, including statistical testing and user-defined parameters, the quantitative effects of suggested manipulations cannot be estimated, and success in improving production is not guaranteed. In contrast, the growth-coupled approach in **Chapter 4** linked the desired pathway to growth, ensuring high fluxes through the pathway *in vivo*. If the growth-coupling approach is successful, Flux Balance Analysis (FBA) simulations with growth as an objective can estimate *in vivo* production fluxes. While **Chapter 3** created a tool that predicts targets with minimal human intervention, **Chapter 4** illustrated the "art" of GEM modeling, which requires experience to pose the right questions and perform appropriate simulations to obtain the sought answers. Although ranking pyruvate-releasing reactions for growth-coupled production was straightforward, decisions on gene deletions and the use of a biosensor during laboratory evolution were not. Similarly, calculating maximum yields based on FBA simulations with production as an objective was easy, but identifying the metabolic engineering strategies to achieve these yields required careful, manual examination of flux predictions. Although one of the strains developed in **Chapter 4** achieved 89.3% of the maximum 4-hydroxybenzoate pathway yield from glycerol minimal medium, its growth rate was very low ($0.008 \pm 0.000\ h^{-1}$). Strategies to improve this growth rate could be devised using CFSA to identify reactions with significant flux changes when "normal" glycerol metabolism is substituted by different variants of the shikimate-dependent catabolism. Additionally, a new round of ALE could be employed to enhance the strain's phenotype, likely through changes in regulators not included in GEMs.

A common drawback of **Chapters 3 and 4**, also shared with other GEM-based strain design approaches, is the ample solution space characteristic of GEM simulations [59]. This results in solutions where reactions that are active in the model are inactive *in vivo* and vice versa. In both cases, the use of extended GEMs such as the enzyme-constrained GEM (ecGEM) employed in **Chapter 5**, which considers the limited capacities of the cells to synthesize and store proteins, could improve predictions. These proteomic constraints facilitate more accurate prediction of metabolic fluxes and cellular phenotypes, and consequently the development of better metabolic engineering strategies for improving the performance of microbial cell factories [24]. Besides, similarly to the pathway optimization process, combinations of metabolic engineering targets suggested by GEMs and their interactions can be studied with the aid of DoE and ML [357].

A focus on bioprocesses

In bioprocess modeling, microbial growth is considered within the framework of a reactor where mass transfer and mixing occur [382]. The translation of laboratory processes to commercial-sized volumes is recognized as the major risk in new bioprocess development [25]. Therefore, process modeling often prioritizes scale-up. Computational fluid dynamics is employed to estimate non-ideal mixing conditions in large vessels and provide insights into nutrient gradients that affect cell responses [25]. While studying the detailed mechanisms of these processes is very relevant, **Chapters 5, 6, and 8** underscore the importance of considering environmental conditions during the design of cell factories. **Chapter 5** focused on metabolic changes in *S. cerevisiae* as a function of the reactor operation, **Chapter 6** explored *P. putida*'s response to oxygen and glucose limitation, and **Chapter 8** proposed DoE as a strategy for simultaneous pathway, media, and process optimization (Figure 10.1).

During process optimization, cell growth is often represented using Monod's equation, cell metabolism is simplified to Herbert-Pirt relations, and efforts are made to model the reactor system [58]. In contrast, when GEMs are employed to model cell metabolism, a detailed description of the metabolic network is possible due to the steady-state assumption [59]. In **Chapter 5**, dynamic Flux Balance Analysis (dFBA) was used to simulate cell metabolism in the reactor environment, expanding the reactor model and providing insights into intracellular fluxes. This allowed the consideration of metabolic changes, such as those occurring at different growth rates, during the strain design process. Furthermore, metabolic changes captured by dFBA can serve as a basis for process control and design. For example, Chang et al. used dFBA to compute glucose feeding and dissolved oxygen profiles that maximize ethanol production in *S. cerevisiae* [383]. Similarly, Raj et al. developed the mcPECASO framework for designing two-stage fermentation processes with optimal titer, rate, and yields based on dFBA simulations [160]. While detailed kinetic models can also be embedded within reactor simulations, they are often limited to a few relevant metabolic pathways [384].

The combination of (ec)GEMs and dFBA in **Chapter 5** allowed for the comparison of yields and productivities among different strains and dynamic production processes. However, various process parameters such as pH, temperature, pressure, or shear stress may affect cell factory performance through mechanisms not included in GEMs [25]. In **Chapter 8**, we proposed estimating the effect of these factors using DoE. Although DoE has traditionally been used for the optimization of the fermentation process [79, 80, 81, 82, 83, 84, 85], this work, along with Zhou et al. and Brown et al., emphasizes the importance of simultaneous strain and bioprocess optimization to avoid selecting sub-optimal strains that only perform well in laboratory settings [88, 89]. In this way, we showed how DoE can identify genetic and process parameters with a significant influence on production, establishing factors that should be prioritized in down/up-scaling plans. Similar to pathway optimization strategies, ML can also be used to study the relationships between relevant environmental factors [94]. However, the throughput of testing multiple production conditions in down-scaling bioreactors is still often limited [50].

While in **Chapter 5**, the study of cell metabolism was limited by available knowledge, the approach described in **Chapter 8** can identify factors that impact the cell's performance without providing explanations about their relevance. Instead, **Chapter 6** aimed to increase understanding regarding how *P. putida* adapts to low oxygen concentrations, commonly encountered in industrial-scale reactors [25, 385]. In this chapter, the use of omics data enabled a comprehensive overview of the cell at the transcriptomic and proteomic scales. Although we only observed a limited change in the proteome of *P. putida* cells grown under oxygen limitation, condition-specific GEMs, constrained with proteomic datasets, can be created when significant metabolic rearrangements are found [59, 381]. Moreover, **Chapter 6** also emphasized the influence of growth conditions on cell physiology. While exponentially growing *P. putida* cells only produce pyoverdine when iron is limiting [282, 313, 320], slow-growing cells produce this compound during glucose-limited growth. Pyoverdine production impacted the biomass yield on glucose and further highlights the interactions between pathways, metabolism, and bioprocess, and the importance of their combinatorial optimization.

In silico DBTL

The throughput for constructing and testing strains is rapidly increasing, enabling the generation of extensive datasets for model training [49]. However, not all research groups have access to such facilities, and the creation of full factorial libraries that test all possible factor combinations becomes rare as the number of factors increases. The lack of a known "ground truth" (*i.e.* the real relationships between factors and their interactions with the response) complicates the comparison between alternative modeling approaches that aim to optimize production. In **Chapters 8 and 9**, we identified the best combinations of tested factor levels for p-coumaric acid production. However, only the *in silico* approach in **Chapter 7** ensured an unbiased evaluation of computational methods. This chapter used one of the kinetic models developed in **Chapter 2** to generate a full factorial library of strains with different concentrations of the pathway enzymes and determine

the actual best combination of factors for optimal production. Subsequently, DoE designs were evaluated based on their ability to identify the true best strains. Similar approaches have been employed to compare methods for pathway optimization, including different ML models [333], reinforcement learning strategies [334], or random sampling and D-optimal designs [335]. Additionally, kinetic models have been used to compare optimal experimental design methods for parameter estimation [386]. As demonstrated in **Chapter 7**, the use of these *in silico* approaches allows the evaluation of the robustness of the computational methods to characteristics inherent to biological datasets, such as missing data or noise. However, although noise is often included in these studies, evaluating realistic scenarios that consider the inability to construct some of the desired strains, as experienced in **Chapters 8 and 9**, should also become a common practice in *in silico* studies. Similarly, including the burden of high expression of multiple genes on cell physiology would also contribute to *in silico* DBTL cycles that more closely resemble their *in vivo* counterparts.

Finally, I would like to highlight the additional educational value of *in silico* studies. Numerous experiments can be easily simulated and multiple hypotheses can be tested and compared enhancing the learning experience and facilitating the understanding of the studied methods.

Optimization and understanding: two sides of the same coin

In recent decades, the field of biology has transitioned from a primarily descriptive science to an engineering discipline [387]. According to ChatGPT, a scientist is "an individual who engages in the systematic and empirical study of the natural world seeking to understand, explain, and predict natural phenomena and explore the underlying principles that govern the universe". In turn, an engineer is "a professional who applies scientific and mathematical principles to design, develop, and create practical solutions to real-world problems". In other words, a scientist aims to understand the natural world while an engineer tries to apply scientific knowledge. In my opinion, a biotechnologist combines traits from both definitions as someone who aspires to comprehend biological principles (understanding) so they can be applied to address relevant problems (optimization).

When the understanding of metabolism was limited, and targeted genetic modification tools were unavailable, random mutagenesis facilitated the over-production of important compounds such as penicillin [27]. Now the knowledge about cell metabolism and molecular biology techniques for directed genetic modification has significantly increased. For example, the construction of the Keio collection of *E. coli* single knock-outs has enabled the systematic analysis of gene function, providing deeper insights into cell behavior [388]. This enhanced understanding has, in turn, driven the rational design of multiple metabolic engineering strategies for metabolite overproduction [174, 180, 186, 188]. Additionally, this knowledge has been gathered in mechanistic models that can also guide the development of new insights. For example, the use of ecGEM and dFBA in **Chapter 5** led to the hypothesis that carbon catabolite repression is essential to obtain maximum growth rates when limitations at the proteome level are considered. A similar conclusion

was recently reached by Liu et al., who based on ecGEM simulations followed by omic analysis, suggest the minimization of proteome reallocation to explain metabolic transitions in sequential utilization of mixed carbon sources of lactic acid bacteria [389]. Likewise, Elsemman et al. developed a GEM with compartment-dependent proteome constraints that provided deeper insights into the physiology of *S. cerevisiae* [170], and Mishra et al. developed a kinetic model of lipid metabolism revealing the presence of a futile cycle in triacylglycerol biosynthesis [118]. These examples show how understanding, especially when collected in knowledge-based models, leads to optimization, the identification of knowledge gaps, and, in turn, new mechanisms. However, when model predictions fail, as experienced in **Chapter 2**, human creativity (and luck) are required to rectify the model, allowing for the generation of new knowledge at the expense of trial-and-error experimental approaches [52].

Rational engineering and mechanistic models are constrained by existing knowledge and result in extended developmental times. For example, compared to the detailed understanding of cell metabolism, less is known about how the cell's regulatory network operates. When regulation is limiting, random mutagenesis, typically in the form of ALE, proves extremely useful for gaining new insights and improving production [380]. In this way, ALE is a good example of how optimization can be followed by understanding. In **Chapter 4**, ALE was employed to rewire *P. putida*'s metabolism for a shikimate-derived catabolism (SDC). Whole-genome sequencing of the evolved strains identified *miaA* and *mexT* as key regulators whose deletion allowed increased fluxes through the shikimate pathway. While the primary goal of the ALE experiment was to optimize SDC, this optimization led to new biological knowledge.

Similarly to how ALE is acknowledged by biologists as a valuable knowledge source, data-driven models can complement mechanistic models, unraveling complex biological mechanisms. However, a universal method to enhance the interpretability of black-box models, equivalent to whole-genome sequencing in ALE experiments, is still missing, which hinders the extraction of knowledge from data-driven approaches. Despite a lack of straightforward interpretation, ML models can lead to the generation of new hypotheses. For instance, ML models trained on metabolome concentrations from strains with different enzyme knockouts enabled the identification of candidate genes crucial for metabolic regulation [390]. Furthermore, ML models are used to estimate parameters such as enzyme catalytic constants [391, 392], which can later be incorporated into mechanistic models [393, 394]. The combination of mechanistic and ML methods enhances their interpretability and facilitates the connection between optimization and understanding. For example, Ma et al. developed a "visible" neural network based on gene ontology to predict *S. cerevisiae* growth as a function of its genotype [395]. Taking a different approach, Dugourd et al. used ML to integrate gene expression data and GEMs to identify possible pathways explaining observed phenotypes [38], and Yuan et al. developed graph neural networks with each node representing a molecular species [396]. In this way, data-driven optimization can lead to new metabolic engineering strategies that, when comprehended, can improve mechanistic understanding [52].

Although the field of explainable ML is expanding [102], we should acknowledge that we might never understand how complex ML algorithms make their predictions, breaking the link between optimization and understanding. Despite the discomfort this affirmation causes in the scientific community, I view this development as a necessary step for biotechnology. I believe that, while ML can accelerate the much-needed application of biotechnological research, it will coexist with the advancement of mechanistic models that serve the inherent human desire for understanding.

Scientific value is typically defined by the understanding of genetic or molecular mechanisms. However, these validated mechanisms are often insufficient for making accurate predictions about complex biological systems [52]. Therefore, a deeper understanding of experimental design and optimization strategies is also relevant for the scientific community. For instance, Cambray et al. identified sequence properties with a significant effect on translation efficiency through the use of a full factorial design and appropriate statistical testing [397]. Similarly, Brown et al. identified genotype-genotype and genotype-environment interactions relevant to understanding ethanol production in *S. cerevisiae* in different environmental conditions, thanks to DoE-based experimentation [89]. Thus, studies such as those described in the *in silico* DBTL section are essential for advancing biological knowledge beyond optimizing the production of a target molecule. Considering that artificial neural networks were inspired by real neurons, it is not unreasonable to think that biology can inspire more and better algorithms [52]. Unraveling mechanisms is important, but their identification should not completely overshadow other exciting developments within biotechnology, a field with the potential to harness both understanding and optimization.

The future: automated biofoundries

Biofoundries are integrated molecular biology facilities that include robotic liquid-handling platforms, high-throughput analytical instruments, and the software, personnel, and data management systems required to run the equipment [49]. The goal of biofoundries is to streamline and accelerate the design, construction, and testing of biological systems, such as engineered microorganisms. Examples of biomanufacturing biofoundries include the SYNBIOCHEM [338], the Agile [96], and the MIT Broad biofoundries [339]. While the MIT biofoundry achieved the production of six out of ten target molecules in 90 days [339], SYNBIOCHEM accomplished the production of 17 out of 25 material monomers in 85 days [338]. However, biofoundries alone will unlikely be enough to increase the number of biotechnological products that reach the market. Yet, as long as high throughput design, experimentation, and analysis are combined with people trained and passionate about multidisciplinary collaborations, biofoundries present a unique opportunity for industrial biotechnology.

Biofoundries accelerate the build and test phases of the DBTL cycle, reducing the effort needed to construct and test strains and facilitating targeted strain construction for hypothesis testing. For instance, the CFSA tool developed in **Chapter 3** suggests some metabolic engineering

strategies that lack direct mechanistic reasoning such as the up-regulation of hydroxymethylglutaryl-CoA synthase (HMGS) for lipid synthesis. With traditional strain construction timelines, there is a trade-off between testing novel targets based on model predictions and improving production. If the experimenter aims at improving production, modifying other reactions, with obvious mechanisms that do not require GEM modeling, would likely be prioritized. Alternatively, high-throughput strain construction enables more exploratory experiments, rather than every construct being an attempt to improve titer [88]. Additionally, facilitating experimentation will also allow a better estimation of parameters in kinetic models thanks, for example, to the generation of time-series data at high resolution [54].

Besides enabling the creation of big datasets, automation will likely reduce systematic errors during experimentation, improving reproducibility and the quality of the generated datasets [52]. However, compared to other fields such as language models used, for example, for the development of ChatGPT [398], the amount of data generated with biological systems is relatively small, yet noisy and heterogeneous [52]. When a few factors are considered, automation will allow the generation of full factorial libraries, and optimal strains will be directly identified. However, considering the vast amount of factors that affect strain performance, brute-force approaches will not be enough to optimize production nor understand mechanisms, and efficient experimental design and data analysis remain crucial [387]. Experimental design might involve the use of kinetic models or GEM to find metabolic engineering targets (**Chapters 2, 4**), the use of statistical methods or ML (**Chapters 8, 9**), or the combination of multiple approaches. For example, Young et al. used GEM simulations to design alternative pathways for itaconic acid production followed by optimization using response surface methods based on I-optimal designs, Plackett Burman designs, and full factorial libraries [87]. However, efficient design and model construction are not enough when the effect of environmental conditions is ignored. For instance, although Khamwachirapithak et al. trained a random forest-based ML algorithm with ethanol production data obtained at 30°C, the system behavior could not be extrapolated at higher temperatures [399]. Therefore, as exemplified in **Chapter 8** efficient experimental designs that consider relevant environmental factors are crucial. Additionally, as shown in **Chapter 6**, testing strains in multiple conditions can improve our knowledge regarding how microorganisms adapt to living in bioreactors. In turn, this could result in new process designs, an area where innovation is slow compared to the rapid advance of genetic and metabolic engineering [243]. As exemplified in **Chapter 7**, information gain from full factorial libraries is minimal compared to efficient designs, and I believe high throughput is better used when testing more factors than when testing everything.

While experimental design and mathematical modeling can link the design and learning phases of DBTL cycles, new designs based on knowledge from previous rounds must be implemented during building. This implementation is not straightforward: a GEM simulation might indicate the optimal flux through a reaction, a kinetic model can suggest an optimal enzyme concentration or a data-driven model might advise doubling the strength of a promoter. Accurately implementing any of these recommendations *in vivo* is currently impossible. Even when effort

is taken to characterize genetic parts [44, 45], expression is dependent on factors such as the downstream gene [370] or the gene order in an operon [32] and cannot be accurately predicted. Although modeling tools to predict expression levels from sequence data are available [43, 400, 401], they are not perfect, and the relationship between gene expression, enzyme concentration, fluxes, and metabolites increases the complexity of the problem. Alternatively, as shown in **Chapter 9**, genetic factors can be considered categorical features, limiting the need for characterization. This simplification, however, misses some of the quantitative information about these variables: if gene expression is continuous, an optimum can be found, which is not possible if it is represented by categories. This can affect metabolic optimization when only the optimal flux through a reaction, like the HMGS example in **Chapter 3**, improves production. The need for characterization should be coupled with the need for standardization creating standardized, open source, protocols for building and testing as well as adhering to FAIR (Findability, Accessibility, Interoperability, and Reusability) principles during data and metadata generation [26, 402, 403].

Automation can convert biofoundries into self-driving labs (SDL) that combine robotics for automated experiments and data collection with artificial intelligence (AI) that uses these data to recommend follow-up experiments [404]. Casini et al. and Robinson et al. highlight how different optimization strategies are required for different target molecules and how the combination of design algorithms, bioinformatics, manual curation, and literature search is needed [338, 339]. In this context, the optimization strategies described in all chapters of this thesis endow the scientist with information to make decisions in an unbiased and informed way during sequential DBTL cycles. Especially, the methods presented in **Chapters 8 and 9** efficiently suggest the performance of new experiments based on the coefficients of linear models or feature importance and SHAP values. Similarly, CFSA (**Chapter 3**) directly provides metabolic engineering targets for implementation. The next step to achieve an SDL is the use of ML algorithms able to autonomously design sequential experimental rounds based on previous results. For instance, Zhang et al. optimized tryptophan production in yeast [357]. Although the first DBTL cycle required the identification of factors and levels based on GEM simulations and human knowledge, ART, an ML algorithm, suggested levels to test in subsequent rounds to balance exploration of the design space and the completion of superior strains. Similarly, Pandi et al. developed METIS, a versatile active ML workflow with a simple online interface for the data-driven optimization of biological targets with minimal experiments [100]. This algorithm designs random factor combinations for the first DBTL or leverages preexisting datasets and uses active learning to suggest posterior DBTL runs. METIS also includes a feature importance module to identify the most crucial components during system optimization, providing the basis for a deeper understanding of the system itself. When these types of algorithms are connected to robotic systems, DBTL cycles can be performed without the need for human intervention [98].

Beyond the combination of laboratory automation and ML, automated scientists are possible [405]. For instance, Lila is a system developed by Amyris that autonomously made more than 100,000 *in silico* designs targeting the production of 454 small molecules, ordered 1,850 genes for synthesis, created 32,000 distinct microbial strains and analyzed more than 10,000,000 data points

throughout 105 partially overlapping DBTL cycles [405]. Lila rapidly generates metabolic routes, identifies genetic elements for perturbation, and specifies the design and re-design of microbial strains within seconds to minutes. The strains outlined by Lila are then constructed and phenotyped as part of a largely automated in-house pipeline.

Although human intervention is not required, scientists play a role in curating choices made by the system, such as refining the metabolic model or suggesting custom protein modifications [405]. An overview of Lila's modules and their relationship with some of the methods in this thesis is presented in Figure 10.2. Although AI has the potential to revolutionize biology, its implementation is currently extremely expensive [406]. Besides, the use of AI comes with a set of ironies including how our capacity to understand it and adapt to its limitations and biases decreases as its intelligence increases [407]. To avoid deskilling, AI should be able to work with people and facilitate, but not always replace, human decision-making as a key objective in their design [407].

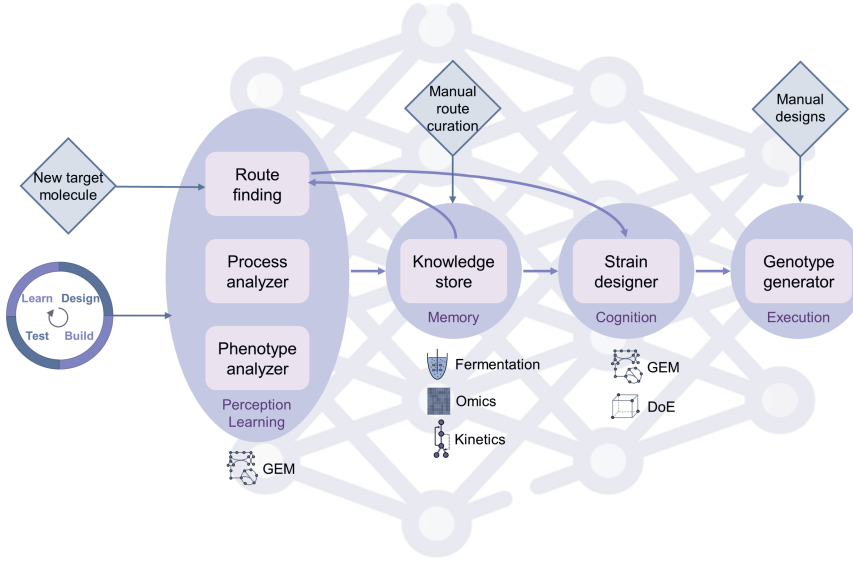


Figure 10.2: Example of an automated scientist. Parts of the Lila platform are analogous to a human scientist's workflow. Diamonds represent input points in which human scientists interact with Lila. The use of fermentation, omics, and kinetics data as well as the use of genome scale metabolic modeling (GEM) and design of experiments (DoE) is highlighted. This figure is a modification of Figure 2 from Singh et al. [405].

It is not clear whether the advantages of biofoundries and SDLs stem from enhanced algorithms or simply increased testing capacities and data availability. From a practical standpoint, achieving economic feasibility in biotechnological processes, whether through improved algorithms or increased throughput, is what matters. However, the journey to achieve this goal is important from a scientific perspective, as it will unveil new exciting knowledge on how numerous genetic and environmental factors affect cell physiology and, thus, the performance of cell factories. These lessons will most likely be only the tip of the iceberg, and new unpredictable discoveries are likely

to occur along the way. Moreover, when multiple factors are considered, no throughput will be sufficient, and truly AI-driven methods (and their associated technology) will emerge. Therefore, the scientific community must have access to biofoundries, and initiatives like the Global Biofoundry Alliance are crucial. In this alliance academic biofoundries collaborate to facilitate the collective sharing of experiences and resources, working together to overcome shared challenges and unmet scientific and engineering needs [408].

Besides the technical difficulties, one of the main challenges that hinders the development of SDLs is, in my opinion, sociological: computer scientists and automation engineers need to work together with molecular and synthetic biologists [387, 409]. These communities do not only differ on their problem-solving approach, but, more importantly on which problems they consider worth solving [52, 404]. Biologists do not need to turn into mathematicians, nor do mathematicians into biologists. However, effective communication between these groups must be achieved. Therefore, an education that includes multidisciplinary projects to favor collaboration, as well as fosters personal development, for instance, through Inner Development Goals, is essential [410]. Open-minded people willing to recognize the limit of their knowledge, eager to share the personal motivations behind their work, and prone to listen to those with a different background and appreciate their efforts, will therefore become key players in the new development of biotechnology.

Can biotechnology fulfill its promises?

The work presented in this thesis contributes to the creation of model-driven approaches applied in DBTL cycles for the design of microbial cell factories and bioprocesses, thereby enhancing our understanding of biological systems. When modeling approaches are coupled with laboratory automation and AI for the creation of self-driven labs, the potential for experimentation becomes unprecedented. This potential will be harnessed for the successful development of biotechnological processes as well as to facilitate the validation of fundamental research.

In July 2023, over three-quarters of the European Union citizens considered climate change a very serious problem [411]. This is not surprising: when one opens any news provider they are likely to find a permanent section on wildfires, droughts, heatwaves, storms, or flooding. Rising greenhouse gas concentrations underscore the urgency of climate action, making it a dominant global risk in the coming decade [412]. Encouragingly, despite its controversy, the 28th United Nations Climate Change Conference (COP28) in December 2023 acknowledged the need to transition away from fossil fuels [18]. This presents an opportunity for biotechnology to fulfill its promises and offer sustainable alternatives to hundreds of petroleum-based chemicals.

However, for biotechnology to play a significant role in the fight against climate change, we must not forget that bio-based and sustainable are not synonyms. While we cannot ignore the fact that the bioeconomy is embedded in the socioeconomic context and needs to create value and compete with traditional industries [12], emphasis must be placed on the use of sustainable substrates and life cycle assessments for bio-based products [17, 26, 413]. Aligning with the plane-

tary boundaries framework, which aims to establish a safe operating space for humanity within the Earth's environmental limits, one of the few powerful means of combating climate change is restoring global forest cover to late 20th century levels [414]. However, the current focus on biomass as a replacement for fossil fuels may hinder this reforestation effort [414]. Hence, there is a pressing need for alternative substrates, focusing on waste degradation, such as the use of lignocellulosic biomass [187, 415, 416], and the capture and conversion of C1 gases [211, 417, 418, 419]. Although biotechnology is now centered on the production of high-value-added products, establishing bio-based production of bulk chemicals and fuels will have a more significant impact on reducing greenhouse gas emissions [14, 24]. In all these applications, life cycle assessments that consider the recycling and biodegradability of bio-based products are essential [17], and legislative efforts that fight "greenwashing" are facilitating steps in the right direction [420].

We are facing unprecedented challenges that demand exciting scientific and technological solutions. Tackling climate change collectively is imperative, and even small contributions must be celebrated. As demonstrated in this thesis, efficient links between all DBTL phases can advance biotechnology, leading to improved cell factories and new biological knowledge that, in turn, support the creation of a sustainable, bio-based future.

Summary

Biotechnology harnesses the power of nature and translates it into applications that improve the quality of our lives and the environment. These applications range from food production through fermentation, the discovery of new drugs, or the development of crops resistant to pests, to production of biofuels or cultivated meat. Specifically, industrial biotechnology focuses on the use of microorganisms, referred to as cell factories, to produce bio-based chemicals as an alternative to petroleum-based production contributing to the fight against climate change. However, designing these cell factories is difficult since we do not completely understand how a cell works and how the desired production affects the functioning of the cell. Moreover, designing bio-process (*i.e.* how the cell will grow and produce at scale) is hindered by the lack of predictability of the performance of the cell factory in industrial settings compared to laboratory scale. The complex interplay among the numerous factors that affect the performance of cell factories prevents accurate predictions of how the microorganism will behave when genetic or environmental variables are perturbed. Instead, bioprocess engineers often have to go through intensive experimentation to optimize production resulting in long developmental times that make the market implementation of biotechnological processes difficult.

Design-Build-Test-Learn (DBTL) cycles are a systematic approach to strain and bioprocess design to iteratively improve the performance of a biological system. In these cycles cells and experiments are designed to acquire the desired information gain, strains are built and tested in the specified conditions, and data is gathered and analyzed to inform the design phase of the following cycle. To leverage all the information acquired in DBTL cycles, and accelerate the design of cell factories, the design and learning phases must be efficiently linked. This can be achieved using mathematical modeling, which encompasses knowledge-based and data-driven models. In this thesis, knowledge-based models including kinetic models (**Chapter 2**) and genome-scale metabolic modeling (**Chapters 3, 4, 5**) are combined with the analysis of omics data (**Chapter 6**) and the use of statistical design of experiments (**Chapter 7, 8**) and machine learning (**Chapter 9**).

Chapters 2 to 4 use mechanistic models to optimize the production of curcumin, a molecule with application in the food and pharmaceutical industries, to design metabolic engineering strategies for growth-uncoupled production of target metabolites, and to guide the construction of *Pseudomonas putida* strains with a new-to-nature carbon catabolism that improves production of shikimate-derived metabolites. In **Chapter 2**, we employed dynamic pathway modeling,

systematic testing of isoenzymes, and optimization of gene expression levels and substrate concentrations in *P. putida* to enhance the biosynthesis of curcuminoids. The use of kinetic ensemble models guided the design of production strains, emphasizing the importance of tuning gene expression. The optimized strain achieved 10.8 $\pm$ 1.8% of the maximum curcumin yield on tyrosine, representing a 4.1-fold increase in production efficiency and the highest reported yield to date. **Chapter 3** introduces Comparative Flux Sampling Analysis (CFSA), a strain design method that involves comparing complete metabolic spaces associated with maximal or near-maximal growth and production phenotypes based on genome-scale metabolic modeling. The comparison, supported by statistical analysis, identifies reactions with altered flux, suggesting targets for genetic interventions such as up-regulations, down-regulations, and gene deletions. CFSA was applied to the production of lipids by *Cutaneotrichosporon oleaginosus* and naringenin by *Saccharomyces cerevisiae*, successfully identifying engineering targets consistent with previous studies and proposing new interventions. **Chapter 4** employs, in turn, a combination of metabolic modeling, rational engineering, and adaptive laboratory evolution to radically refactor bacterial metabolism. Specifically, a new-to-nature shikimate pathway-dependent catabolism was created in *P. putida* by reprogramming the shikimate pathway as the dominant pathway for growth. Whole-genome sequencing of the evolved strains identified *miaA* and *mexT* as key regulators whose deletion allowed increased fluxes through the shikimate pathway. The resulting strain diverts the majority of its carbon catabolism flux through the shikimate pathway, producing 0.35 mol/mol 4-hydroxybenzoate in glycerol minimal medium during growth, achieving 89.2% of the maximum predicted pathway yield. These chapters prove the potential of knowledge-based models for metabolic and pathway optimization. Additionally, while **Chapters 2 and 4** highlight the versatility of *P. putida*'s metabolism and its potential for the production of complex compounds, **Chapter 3** provides a robust, easy-to-use, host-independent method for the design of metabolic engineering strategies.

In **Chapters 5 and 6** the focus changes to understanding how cells adapt to living in bioreactors at the transcriptional, proteome, and metabolic levels. In **Chapter 5** we used Flux balance analysis (FBA) and dynamic FBA (dFBA) to predict the growth dynamics of *S. cerevisiae* under various industrially relevant conditions using a genome-scale model (GEM) and its enzyme-constrained version (ecGEM). The ecGEM outperformed the GEM and, in combination with dFBA and flux sampling, facilitated linking reactor operation and genetic modifications to flux predictions. This enabled the prediction of yields and productivities for different strains and dynamic production processes. Besides, the proposed approach suggested a role of proteome limitation on carbon catabolite repression. **Chapter 6** delved into the response of *P. putida* cells to oxygen and glucose limitations through the use of chemostat cultivations and transcriptomic and proteomic analysis. We report an up to 59% increase in biomass yield of slow-growing cells in oxygen limitation compared to glucose-limited growth due to the absence of pyoverdine production. Our analysis additionally identified 923 differentially expressed genes specific to oxygen-limited growth, with only seven differentially abundant proteins, suggesting *P. putida*'s resilience to long-term oxygen-limited growth. Both of these chapters highlight the effect of growth conditions on cell physiology and the importance

of considering bio-process-related factors during strain and bio-process design.

In **Chapters 7 to 9** I move towards the application of data-driven models to optimize cell factories and understand the relationships between factors that affect the performance of the strains. **Chapter 7** presents a theoretical study on the use of Design of Experiments (DoE) for pathway optimization. Leveraging one of the kinetic models developed in **Chapter 2**, the performance of a full factorial strain library is compared to resolution V, IV, III, and Plackett Burman designs. Assessing robustness to noise and missing data, we suggest using resolution IV designs for the optimization of the expression of pathway genes. These designs enable the identification of optimal strains and provide valuable information regarding the impact of factors and their interactions on production, offering guidance for subsequent optimization cycles. Lessons learned in **Chapter 7** were applied in **Chapter 8** where we used DoE to systematically explore the relationships between media, process, and genetic factors and optimized the production of p-coumaric acid (pCA), a precursor for a wide array of biologically relevant molecules, in *S. cerevisiae*. Two rounds of fractional factorial designs identified factors significantly affecting pCA production, resulting in a 168-fold variation in pCA titer. Additionally, the study revealed a significant interaction between culture temperature and the expression of ARO4, emphasizing the importance of simultaneous process and strain optimization, already highlighted in **Chapters 5 and 6**. In **Chapter 9** we maintain the aim to improve pCA production in *S. cerevisiae* but, this time, we advocate for the generation of a random library of strains with different genes and expression strengths and its analysis using machine learning (ML). The use of a screening before sequencing approach allowed stratification during training and improved ML performance on small datasets. Besides, explainable ML techniques were employed to guide the expansion of the original design space. This approach ultimately led to a 68% increased production of pCA within two DBTL cycles. These chapters underscore the potential of data-driven models to link the design and learning phases of sequential DBTL cycles and facilitate the design of strains and bio-processes.

This dissertation ends with a general discussion (**Chapter 10**) that delves into the application of the explored modeling approaches at different stages during the development of strains and bioprocesses, including pathway, metabolic, and process design. I further discuss the interconnection between understanding and optimization in biotechnological research and the impact of automated biofoundries and modeling on industrial biotechnology. In addition, I reflect on the need to use sustainable substrates and life cycle assessments to evaluate the sustainability of bio-based products.

References

[1] *The White House. "Executive Order on Advancing Biotechnology and Biomanufacturing Innovation for a Sustainable, Safe, and Secure American Bioeconomy"*. 2022. URL: https://www.whitehouse.gov/briefing-room/presidential-actions/2022/09/12/executive-order-on-advancing-biotechnology-and-biomanufacturing-innovation-for-a-sustainable-safe-and-secure-american-bioeconomy/.

[2] Erickson, B., Nelson, J. E., and Winters, P. "Perspective on opportunities in industrial biotechnology in renewable chemicals". In: *Biotechnology Journal* 7.2 (2012), 176–185. DOI: 10.1002/BIOT.201100069.

[3] Schürrle, K. "History, current state, and emerging applications of industrial biotechnology". In: *Advances in Biochemical Engineering/Biotechnology* 173 (2020), 13–51. DOI: 10.1007/10_2018_81.

[4] Aguilar, A., Twardowski, T., and Wohlgemuth, R. "Bioeconomy for Sustainable Development". In: *Biotechnology Journal* 14.8 (2019), 1800638. DOI: 10.1002/BIOT.201800638.

[5] Cappell, K. M. and Kochenderfer, J. N. "Long-term outcomes following CAR T cell therapy: what we know so far". In: *Nature Reviews Clinical Oncology* 20.6 (2023), 359–371. DOI: 10.1038/s41571-023-00754-1.

[6] Lewis, L. M., Badkar, A. V., Cirelli, D., Combs, R., and Lerch, T. F. "The Race to Develop the Pfizer-BioNTech COVID-19 Vaccine: From the Pharmaceutical Scientists' Perspective". In: *Journal of Pharmaceutical Sciences* 112.3 (2023), 640–647. DOI: 10.1016/J.XPHS.2022.09.014.

[7] Dijk, M. van, Morley, T., Rau, M. L., and Saghai, Y. "A meta-analysis of projected global food demand and population at risk of hunger for the period 2010–2050". In: *Nature Food* 2 (2021), 494–501. DOI: 10.1038/s43016-021-00322-9.

[8] Francis, D., Finer, J. J., and Grotewold, E. "Challenges and opportunities for improving food quality and nutrition through plant biotechnology". In: *Current Opinion in Biotechnology* 44 (2017), 124–129. DOI: 10.1016/J.COPBIO.2016.11.009.

[9] Woo, S. L. and Pepe, O. "Microbial consortia: Promising probiotics as plant biostimulants for sustainable agriculture". In: *Frontiers in Plant Science* 9 (2018), 428205. DOI: 10.3389/FPLS.2018.01801.

[10] *European Comission. Closing the loop - An EU action plan for the Circular Economy*. Tech. rep. 2015.

[11] Bell, J., Paula, L., Dodd, T., Németh, S., Nanou, C., Mega, V., and Campos, P. "EU ambition to build the world's leading bioeconomy—Uncertain times demand innovative and sustainable solutions". In: *New Biotechnology* 40 (2018), 25–30. DOI: 10.1016/J.NBT.2017.06.010.

[12] *Agency of European Environment. Innovating for Sustainable Growth: a Bioeconomy for Europe, EC, 2012 — European Environment Agency*. 2012. URL: https://www.eea.europa.eu/policy-documents/innovating-for-sustainable-growth-a.

[13] *Acumen Public Affairs. Biotechnology: Is the EU about to up its game?* URL: https://acumenpa.com/biotechnology-is-the-eu-about-to-up-its-game/.

[14] Lee, J. A., Kim, H. U., Na, J.-G., Ko, Y.-S., Cho, J. S., and Lee, S. Y. "Factors affecting the competitiveness of bacterial fermentation". In: *Trends in Biotechnology* 41.6 (2023). DOI: 10.1016/j.tibtech.2022.10.005.

[15] Nielsen, J., Tillegreen, C. B., and Petranovic, D. "Innovation trends in industrial biotechnology". In: *Trends in Biotechnology* 40.10 (2022), 1160–1172. DOI: 10.1016/J.TIBTECH.2022.03.007.

[16] Bugge, M. M., Hansen, T., and Klitkou, A. "What is the bioeconomy?" In: *From Waste to Value: Valorisation Pathways for Organic Waste Streams in Circular Bioeconomies*. Taylor and Francis, 2019, 19–50. ISBN: 9780429863257. DOI: 10.4324/9780429460289-2.

[17] Stegmann, P., Londo, M., and Junginger, M. "The circular bioeconomy: Its elements and role in European bioeconomy clusters". In: *Resources, Conservation & Recycling* 6 (2020), 100029. DOI: 10.1016/J.RCRX.2019.100029.

[18] *COP28 UAE - United Nations Climate Change Conference*. URL: https://www.cop28.com/en/.

[19] Philp, J. "The bioeconomy, the challenge of the century for policy makers". In: *New Biotechnology* 40 (2018), 11–19. DOI: 10.1016/J.NBT.2017.04.004.

[20] Brossard, D. "Biotechnology, communication and the public: Keys to delve into the social perception of science". In: *Metode Science Studies Journal* 0.9 (2019), 39–45. DOI: 10.7203/METODE.9.11347.

[21] Woźniak, E., Tyczewska, A., and Twardowski, T. "A Shift Towards Biotechnology: Social Opinion in the EU". In: *Trends in Biotechnology* 39.3 (2021), 214–218. DOI: 10.1016/j.tibtech.2020.08.001.

[22] *THE 17 GOALS | Sustainable Development*. URL: https://sdgs.un.org/goals.

[23] El-Chichakli, B., Von Braun, J., Lang, C., Barben, D., and Philp, J. "Policy: Five cornerstones of a global bioeconomy". In: *Nature* 535.7611 (2016), 221–223. DOI: 10.1038/535221a.

[24] Kim, G. B., Choi, S. Y., Cho, I. J., Ahn, D. H., and Lee, S. Y. "Metabolic engineering for sustainability and health". In: *Trends in Biotechnology* 41.3 (2023), 425–451. DOI: 10.1016/J.TIBTECH.2022.12.014.

[25] Cordell, W. T., Avolio, G., Takors, R., and Pfleger, B. F. "Milligrams to kilograms: making microbes work at scale". In: *Trends in Biotechnology* 41.11 (2023), 1442–1457. DOI: 10.1016/J.TIBTECH.2023.05.002.

[26] Hodgson, A., Maxon, M. E., and Alper, J. "The U.S. Bioeconomy: Charting a Course for a Resilient and Competitive Future". In: *Industrial Biotechnology* 18.3 (2022), 115–136. DOI: 10.1089/IND.2022.29283.AHO.

[27] Thykaer, J. and Nielsen, J. "Metabolic engineering of β-lactam production". In: *Metabolic Engineering* 5.1 (2003), 56–69. DOI: 10.1016/S1096-7176(03)00003-X.

[28] Hodgman, C. E. and Jewett, M. C. "Cell-free synthetic biology: Thinking outside the cell". In: *Metabolic Engineering* 14.3 (2012), 261–269. DOI: 10.1016/J.YMBEN.2011.09.002.

[29] Kwok, R. "Five hard truths for synthetic biology: can engineering approaches tame the complexity of living systems?" In: *Nature* 463.7279 (2010), 288–291.

[30] Kampers, L. F., Asin-Garcia, E., Schaap, P. J., Wagemakers, A., and Martins dos Santos, V. A. "Navigating the Valley of Death: Perceptions of Industry and Academia on Production Platforms and Opportunities in Biotechnology". In: *EFB Bioeconomy Journal* 2 (2022), 100033. DOI: 10.1016/J.BIOECO.2022.100033.

[31] Gurdo, N., Volke, D. C., McCloskey, D., and Nikel, P. I. "Automating the design-build-test-learn cycle towards next-generation bacterial cell factories". In: *New Biotechnology* 74 (2023), 1–15. DOI: 10.1016/J.NBT.2023.01.002.

[32] Carbonell, P. et al. "An automated Design-Build-Test-Learn pipeline for enhanced microbial production of fine chemicals". In: *Communications Biology* 1.1 (2018), 66. DOI: 10.1038/s42003-018-0076-9.

[33] Thomas, S., Maynard, N. D., and Gill, J. "DNA library construction using Gibson Assembly®". In: *Nature Methods* 12.11 (2015), i–ii. DOI: 10.1038/nmeth.f.384.

[34] Durante-Rodríguez, G., De Lorenzo, V., and Martínez-García, E. "The standard European vector architecture (SEVA) plasmid toolkit". In: *Methods in Molecular Biology* 1149 (2014), 469–478. DOI: 10.1007/978-1-4939-0473-0_36.

[35] Wang, H., La Russa, M., and Qi, L. S. "CRISPR/Cas9 in Genome Editing and Beyond". In: *Annual Review of Biochemistry* 85 (2016), 227–264. DOI: 10.1146/ANNUREV-BIOCHEM-060815-014607.

[36] Becker, J. and Wittmann, C. "From systems biology to metabolically engineered cells — an omics perspective on the development of industrial microbes". In: *Current Opinion in Microbiology* 45 (2018), 180–188. DOI: 10.1016/J.MIB.2018.06.001.

[37] Badia-I-Mompel, P., Vélez Santiago, J., Braunger, J., Geiss, C., Dimitrov, D., Müller-Dott, S., Taus, P., Dugourd, A., Holland, C. H., Ramirez Flores, R. O., and Saez-Rodriguez, J. "decoupleR: ensemble of computational methods to infer biological activities from omics data". In: *Bioinformatics Advances* 2.1 (2022). DOI: 10.1093/BIOADV/VBAC016.

[38] Dugourd, A. et al. "Causal integration of multi-omics data with prior knowledge to generate mechanistic hypotheses". In: *Molecular Systems Biology* 17.1 (2021), e9730. DOI: 10.15252/MSB.20209730.

[39] Wang, X., Zhao, Y., Hou, Z., Chen, X., Jiang, S., Liu, W., Hu, X., Dai, J., and Zhao, G. "Large-scale pathway reconstruction and colorimetric screening accelerate cellular metabolism engineering". In: *Metabolic Engineering* 80 (2023), 107–118. DOI: 10.1016/J.YMBEN.2023.09.009.

[40] Van Brempt, M., Peeters, A. I., Duchi, D., De Wannemaeker, L., Maertens, J., De Paepe, B., and De Mey, M. "Biosensor-driven, model-based optimization of the orthogonally expressed naringenin biosynthesis pathway". In: *Microbial Cell Factories* 21.1 (2022), 1–19. DOI: 10.1186/S12934-022-01775-8.

[41] Rohe, P., Venkanna, D., Kleine, B., Freudl, R., and Oldiges, M. "An automated workflow for enhancing microbial bioprocess optimization on a novel microbioreactor platform". In: *Microbial Cell Factories* 11.1 (2012), 1–14. DOI: 10.1186/1475-2859-11-144.

[42] Janakiraman, V., Kwiatkowski, C., Kshirsagar, R., Ryll, T., and Huang, Y. M. "Application of high-throughput mini-bioreactor system for systematic scale-down modeling, process characterization, and control strategy development". In: *Biotechnology Progress* 31.6 (2015), 1623–1632. DOI: 10.1002/BTPR.2162.

[43] Jeschek, M., Gerngross, D., and Panke, S. "Rationally reduced libraries for combinatorial pathway optimization minimizing experimental effort". In: *Nature Communications* 7.1 (2016), 1–10. DOI: 10.1038/ncomms11163.

[44] Elmore, J. R., Furches, A., Wolff, G. N., Gorday, K., and Guss, A. M. "Development of a high efficiency integration system and promoter library for rapid modification of *Pseudomonas putida* KT2440". In: *Metabolic Engineering Communications* 5 (2017), 1–8. DOI: 10.1016/J.METENO.2017.04.001.

[45] Guo, Y., Dong, J., Zhou, T., Auxillos, J., Li, T., Zhang, W., Wang, L., Shen, Y., Luo, Y., Zheng, Y., Lin, J., Chen, G. Q., Wu, Q., Cai, Y., and Dai, J. "YeastFab: the design and construction of standard biological parts for metabolic engineering in *Saccharomyces cerevisiae*". In: *Nucleic Acids Research* 43.13 (2015), e88–e88. DOI: 10.1093/NAR/GKV464.

[46] Kanehisa, M. and Goto, S. "KEGG: Kyoto Encyclopedia of Genes and Genomes". In: *Nucleic Acids Research* 28.1 (2000), 27–30. DOI: 10.1093/NAR/28.1.27.

[47] Chang, A., Jeske, L., Ulbrich, S., Hofmann, J., Koblitz, J., Schomburg, I., Neumann-Schaal, M., Jahn, D., and Schomburg, D. "BRENDA, the ELIXIR core data resource in 2021: new developments and updates". In: *Nucleic Acids Research* 49.D1 (2021), D498–D508. DOI: 10.1093/NAR/GKAA1025.

[48] Caspi, R. et al. "The MetaCyc database of metabolic pathways and enzymes and the BioCyc collection of Pathway/Genome Databases". In: *Nucleic Acids Research* 42.D1 (2014), D459–D471. DOI: 10.1093/nar/gkt1103.

[49] Tellechea-Luzardo, J., Otero-Muras, I., Goñi-Moreno, A., and Carbonell, P. "Fast biofoundries: coping with the challenges of biomanufacturing". In: *Trends in Biotechnology* 40.7 (2022), 831–842. DOI: 10.1016/J.TIBTECH.2021.12.006.

[50] Liao, X., Ma, H., Tang, Y. J., Xu, P., and Zhou, K. "Artificial intelligence: a solution to involution of design–build–test–learn cycle". In: *Current Opinion in Biotechnology* 75 (2022), 102712. DOI: 10.1016/J.COPBIO.2022.102712.

[51] Otero-Muras, I. and Carbonell, P. "Automated engineering of synthetic metabolic pathways for efficient biomanufacturing". In: *Metabolic Engineering* 63 (2021), 61–80. DOI: 10.1016/J.YMBEN.2020.11.012.

[52] Lawson, C., Martí, J. M., Radivojevic, T., Jonnalagadda, S. V. R., Gentz, R., Hillson, N. J., Peisert, S., Kim, J., Simmons, B. A., Petzold, C. J., Singer, S. W., Mukhopadhyay, A., Tanjore, D., Dunn, J., and Martin, H. G. "Machine learning for metabolic engineering: A review". In: *Metabolic Engineering* (2020). DOI: 10.1016/j.ymben.2020.10.005.

[53] The New Yorker. *How Much of the World Is It Possible to Model?* URL: https://www.newyorker.com/culture/annals-of-inquiry/how-much-of-the-world-is-it-possible-to-model.

[54] Rosmalen, R. P. van, Martins dos Santos, V. A. P., and Suarez-Diez, M. "Questions, data and models underpinning metabolic engineering". In: *Frontiers in Systems Biology* 2 (2022), 998048. DOI: 10.3389/FSYSB.2022.998048.

[55] Saa, P. A. and Nielsen, L. K. "Formulation, construction and analysis of kinetic models of metabolism: A review of modelling frameworks". In: *Biotechnology Advances* 35.8 (2017), 981–1003. DOI: 10.1016/J.BIOTECHADV.2017.09.005.

[56] Koutinas, M., Kiparissides, A., Pistikopoulos, E. N., and Mantalaris, A. "Bioprocess systems engineering: Transferring traditional process engineering principles to industrial biotechnology". In: *Computational and Structural Biotechnology Journal* 3.4 (2012), e201210022. DOI: 10.5936/csbj.201210022.

[57] Fang, X., Lloyd, C. J., and Palsson, B. O. "Reconstructing organisms in silico: genome-scale models and their emerging applications". In: *Nature Reviews Microbiology* 18.12 (2020), 731–743. DOI: 10.1038/s41579-020-00440-4.

[58] Kerssemakers, A. A., Øzmerih, S., Sin, G., and Sudarsan, S. "Dynamic Interplay between O_2 Availability, Growth Rates, and the Transcriptome of *Yarrowia lipolytica*". In: *Fermentation* 9.1 (2023), 74. DOI: 10.3390/FERMENTATION9010074/S1.

[59] Carter, E. L., Constantinidou, C., and Alam, M. T. "Applications of genome-scale metabolic models to investigate microbial metabolic adaptations in response to genetic or environmental perturbations". In: *Briefings in Bioinformatics* 25.1 (2023), 1–14. DOI: 10.1093/BIB/BBAD439.

[60] Foster, C. J., Wang, L., Dinh, H. V., Suthers, P. F., and Maranas, C. D. "Building kinetic models for metabolic engineering". In: *Current Opinion in Biotechnology* 67 (2021), 35–41. DOI: 10.1016/J.COPBIO.2020.11.010.

[61] Tran, L. M., Rizk, M. L., and Liao, J. C. "Ensemble modeling of metabolic networks". In: *Biophysical Journal* 95.12 (2008), 5606–5617. DOI: 10.1529/biophysj.108.135442.

[62] Orth, J. D., Thiele, I., and Palsson, B. Ø. "What is flux balance analysis?" In: *Nature Biotechnology* 28.3 (2010), 245–248. DOI: 10.1038/nbt.1614.

[63] Schuetz, R., Kuepfer, L., and Sauer, U. "Systematic evaluation of objective functions for predicting intracellular fluxes in *Escherichia coli*". In: *Molecular Systems Biology* 3.119 (2007). DOI: 10.1038/msb4100162.

[64] Herrmann, H. A., Dyson, B. C., Vass, L., Johnson, G. N., and Schwartz, J. M. "Flux sampling is a powerful tool to study metabolism under changing environmental conditions". In: *npj Systems Biology and Applications* 5.1 (2019), 1–8. DOI: 10.1038/s41540-019-0109-0.

[65] Henry, C. S., Broadbelt, L. J., and Hatzimanikatis, V. "Thermodynamics-based metabolic flux analysis". In: *Biophysical Journal* 92.5 (2007), 1792–1805. DOI: 10.1529/biophysj.106.093138.

[66] Grigaitis, P., Olivier, B. G., Fiedler, T., Teusink, B., Kummer, U., and Veith, N. "Protein cost allocation explains metabolic strategies in *Escherichia coli*". In: *Journal of Biotechnology* 327 (2021), 54–63. DOI: 10.1016/j.jbiotec.2020.11.003.

[67] Sánchez, B. J., Zhang, C., Nilsson, A., Lahtvee, P.-J., Kerkhoven, E. J., and Nielsen, J. "Improving the phenotype predictions of a yeast genome-scale metabolic model by incorporating enzymatic constraints". In: *Molecular Systems Biology* 13 (2017), 935. DOI: 10.15252/msb.20167411.

[68] Carrera, J., Estrela, R., Luo, J., Rai, N., Tsoukalas, A., and Tagkopoulos, I. "An integrative, multi-scale, genome-wide model reveals the phenotypic landscape of *E scherichia coli*". In: *Molecular Systems Biology* 10.7 (2014). DOI: 10.15252/MSB.20145108.

[69] Banerjee, D., Eng, T., Lau, A. K., Sasaki, Y., Wang, B., Chen, Y., Prahl, J.-P., Singan, V. R., Herbert, R. A., Liu, Y., Tanjore, D., Petzold, C. J., Keasling, J. D., and Mukhopadhyay, A. "Genome-scale metabolic rewiring improves titers rates and yields of the non-native product indigoidine at scale". In: *Nature Communications* 11.1 (2020), 5385. DOI: 10.1038/s41467-020-19171-4.

[70] Øyås, O. and Stelling, J. "Genome-scale metabolic networks in time and space". In: *Current Opinion in Systems Biology* 8 (2018), 51–58. DOI: 10.1016/J.COISB.2017.12.003.

[71] Kumar, V., Bhalla, A., and Rathore, A. S. "Design of experiments applications in bioprocessing: Concepts and approach". In: *Biotechnology Progress* 30.1 (2014), 86–99. DOI: 10.1002/btpr.1821.

[72] Lawson, J. *Design and Analysis of Experiments with R*. Ed. by F. Dominici, J. Faraway, M. Tanner, and J. Zidek. CRC press, 2014. ISBN: 978-1-4987-2848-5.

[73] Gilman, J., Walls, L., Bandiera, L., and Menolascina, F. "Statistical Design of Experiments for Synthetic Biology". In: *ACS Synthetic Biology* 10.1 (2021), 1–18. DOI: 10.1021/acssynbio.0c00385.

[74] Asnicar, F., Thomas, A. M., Passerini, A., Waldron, L., and Segata, N. "Machine learning for microbiologists". In: *Nature Reviews Microbiology* (2023), 1–15. DOI: 10.1038/s41579-023-00984-1.

[75] Volk, M. J., Lourentzou, I., Mishra, S., Vo, L. T., Zhai, C., and Zhao, H. "Biosystems Design by Machine Learning". In: *ACS Synthetic Biology* 9.7 (2020), 1514–1533. DOI: 10.1021/acssynbio.0c00129.

[76] Smucker, B., Krzywinski, M., and Altman, N. "Optimal experimental design". In: *Nature Methods* 15.8 (2018), 559–560. DOI: 10.1038/S41592-018-0083-2.

[77] Rosmalen, R. P. van, Martins dos Santos, V. A., and Suarez-Diez, M. *Model-driven engineering of microbial metabolism*. Wageningen University, 2022, 231. ISBN: 9789464470482.

[78] Ruess, J., Parise, F., Milias-Argeitis, A., Khammash, M., and Lygeros, J. "Iterative experiment design guides the characterization of a light-inducible gene expression circuit". In: *Proceedings of the National Academy of Sciences of the United States of America* 112.26 (2015), 8148–8153. DOI: 10.1073/PNAS.1423947112.

[79] Azubuike, C. C., Edwards, M. G., Gatehouse, A. M., and Howard, T. P. "Applying statistical design of experiments to understanding the effect of growth medium components on *Cupriavidus necator* H16 Growth". In: *Applied and Environmental Microbiology* 86.17 (2020). DOI: 10.1128/AEM.00705-20.

[80] Akbarzadeh, A., Dehnavi, E., Aghaeepoor, M., and Amani, J. "Optimization of Recombinant Expression of Synthetic Bacterial Phytase in *Pichia pastoris* Using Response Surface Methodology". In: *Jundishapur Journal of Microbiology* 8.12 (2015), 27553. DOI: 10.5812/JJM.27553.

[81] Duman-Özdamar, Z. E., Martins dos Santos, V. A., Hugenholtz, J., and Suarez-Diez, M. "Tailoring and optimizing fatty acid production by oleaginous yeasts through the systematic exploration of their physiological fitness". In: *Microbial Cell Factories* 21.1 (2022), 1–13. DOI: 10.1186/S12934-022-01956-5/FIGURES/3.

[82] Lee, Y. J., Kim, H. J., Gao, W., Chung, C. H., and Lee, J. W. "Statistical optimization for production of carboxymethylcellulase of *Bacillus amyloliquefaciens* DL-3 by a recombinant *Escherichia coli* JM109/DL-3 from rice bran using response surface method". In: *Biotechnology and Bioprocess Engineering* 17.2 (2012), 227–235. DOI: 10.1007/S12257-011-0258-5/METRICS.

[83] Motta Dos Santos, L. F., Coutte, F., Ravallec, R., Dhulster, P., Tournier-Couturier, L., and Jacques, P. "An improvement of surfactin production by *B. subtilis* BBG131 using design of experiments in microbioreactors and continuous process in bubbleless membrane bioreactor". In: *Bioresource Technology* 218 (2016), 944–952. DOI: 10.1016/J.BIORTECH.2016.07.053.

[84] Xu, P., Ding, Z. Y., Qian, Z., Zhao, C. X., and Zhang, K. C. "Improved production of mycelial biomass and ganoderic acid by submerged culture of *Ganoderma lucidum* SB97 using complex media". In: *Enzyme and Microbial Technology* 42.4 (2008), 325–331. DOI: 10.1016/j.enzmictec.2007.10.016.

[85] Song, T. Q., Ding, M. Z., Zhai, F., Liu, D., Liu, H., Xiao, W. H., and Yuan, Y. J. "Engineering *Saccharomyces cerevisiae* for geranylgeraniol overproduction by combinatorial design". In: *Scientific Reports* 7.1 (2017), 1–11. DOI: 10.1038/s41598-017-15005-4.

[86] Xu, P., Rizzoni, E. A., Sul, S. Y., and Stephanopoulos, G. "Improving metabolic pathway efficiency by statistical model-based multivariate regulatory metabolic engineering". In: *ACS Synthetic Biology* 6.1 (2017), 148–158. DOI: 10.1021/acssynbio.6b00187.

[87] Young, E. M., Zhao, Z., Gielesen, B. E., Wu, L., Benjamin Gordon, D., Roubos, J. A., and Voigt, C. A. "Iterative algorithm-guided design of massive strain libraries, applied to itaconic acid production in yeast". In: *Metabolic Engineering* 48 (2018), 33–43. DOI: 10.1016/j.ymben.2018.05.002.

[88] Zhou, H., Vonk, B., Roubos, J. A., Bovenberg, R. A., and Voigt, C. A. "Algorithmic co-optimization of genetic constructs and growth conditions: application to 6-ACA, a potential nylon-6 precursor". In: *Nucleic Acids Research* 43.21 (2015). DOI: 10.1093/nar/gkv1071.

[89] Brown, S. R., Staff, M., Lee, R., Love, J., Parker, D. A., Aves, S. J., and Howard, T. P. "Design of Experiments Methodology to Build a Multifactorial Statistical Model Describing the Metabolic Interactions of Alcohol Dehydrogenase Isozymes in the Ethanol Biosynthetic Pathway of the Yeast *Saccharomyces cerevisiae*". In: *ACS Synthetic Biology* 7.7 (2018), 1676–1684. DOI: 10.1021/acssynbio.8b00112.

[90] Costello, Z. and Martin, H. G. "A machine learning approach to predict metabolic pathway dynamics from time-series multiomics data". In: *npj Systems Biology and Applications* 4.1 (2018), 1–14. DOI: 10.1038/s41540-018-0054-3.

[91] Clauwaert, J., Menschaert, G., and Waegeman, W. "DeepRibo: a neural network for precise gene annotation of prokaryotes by combining ribosome profiling signal and binding site patterns". In: *Nucleic Acids Research* 47.6 (2019), e36–e36. DOI: 10.1093/NAR/GKZ061.

[92] Senior, A. W. et al. "Improved protein structure prediction using potentials from deep learning". In: *Nature* 577.7792 (2020), 706–710. DOI: 10.1038/s41586-019-1923-7.

[93] Koch, M., Duigou, T., and Faulon, J. L. "Reinforcement learning for bioretrosynthesis". In: *ACS Synthetic Biology* 9.1 (2020), 157–168. DOI: 10.1021/ACSSYNBIO.9B00447.

[94] Oyetunde, T., Liu, D., Martin, H. G., and Tang, Y. J. "Machine learning framework for assessment of microbial factory performance". In: *PLOS ONE* 14.1 (2019), e0210558. DOI: 10.1371/journal.pone.0210558.

[95] Lee, M. E., Aswani, A., Han, A. S., Tomlin, C. J., and Dueber, J. E. "Expression-level optimization of a multi-enzyme pathway in the absence of a high-throughput assay". In: *Nucleic Acids Research* 41.22 (2013), 10668–10678. DOI: 10.1093/nar/gkt809.

[96] Opgenorth, P. et al. "Lessons from Two Design-Build-Test-Learn Cycles of Dodecanol Production in *Escherichia coli* Aided by Machine Learning". In: *ACS Synthetic Biology* (2019). DOI: 10.1021/acssynbio.9b00020.

[97] Zhou, Y., Li, G., Dong, J., Xing, X.-h., Dai, J., and Zhang, C. "MiYA, an efficient machine-learning workflow in conjunction with the YeastFab assembly strategy for combinatorial optimization of heterologous metabolic pathways in *Saccharomyces cerevisiae*". In: *Metabolic Engineering* 47 (2018), 294–302. DOI: 10.1016/J.YMBEN.2018.03.020.

[98] HamediRad, M., Chao, R., Weisberg, S., Lian, J., Sinha, S., and Zhao, H. "Towards a fully automated algorithm driven platform for biosystems design". In: *Nature Communications* 10.1 (2019), 1–10. DOI: 10.1038/s41467.019.13189.z.

[99] Radivojević, T., Costello, Z., Workman, K., and Garcia Martin, H. "A machine learning Automated Recommendation Tool for synthetic biology". In: *Nature Communications* 11.1 (2020), 1–14. DOI: 10.1038/s41467.020.18008.4.

[100] Pandi, A., Diehl, C., Yazdizadeh Kharrazi, A., Scholz, S. A., Bobkova, E., Faure, L., Nattermann, M., Adam, D., Chapin, N., Foroughijabbari, Y., Moritz, C., Paczia, N., Cortina, N. S., Faulon, J. L., and Erb, T. J. "A versatile active learning workflow for optimization of genetic and metabolic networks". In: *Nature Communications* 13.1 (2022), 1–15. DOI: 10.1038/s41467-022-31245-z.

[101] Cao, B., Adutwum, L. A., Oliynyk, A. O., Luber, E. J., Olsen, B. C., Mar, A., and Buriak, J. M. "How to optimize materials and devices via design of experiments and machine learning: Demonstration using organic photovoltaics". In: *ACS Nano* 12.8 (2018), 7434–7444. DOI: 10.1021/ACSNANO.8B04726.

[102] Belle, V. and Papantonis, I. "Principles and Practice of Explainable Machine Learning". In: *Frontiers in Big Data* 4 (2021), 688969. DOI: 10.3389/FDATA.2021.688969.

[103] Novakovsky, G., Dexter, N., Libbrecht, M. W., Wasserman, W. W., and Mostafavi, S. "Obtaining genetics insights from deep learning via explainable artificial intelligence". In: *Nature Reviews Genetics* 24.2 (2023), 125–137. DOI: 10.1038/s41576-022-00532-2.

[104] Kim, I. K., Roldão, A., Siewers, V., and Nielsen, J. "A systems-level approach for metabolic engineering of yeast cell factories". In: *FEMS Yeast Research* 12.2 (2012), 228–248. DOI: 10.1111/J.1567-1364.2011.00779.X.

[105] Martin-Pascual, M., Batianis, C., Bruinsma, L., Asin-Garcia, E., Garcia-Morales, L., Weusthuis, R. A., Kranenburg, R. van, and Martins dos Santos, V. A. "A navigation guide of synthetic biology tools for *Pseudomonas putida*". In: *Biotechnology Advances* 49 (2021), 107732. DOI: 10.1016/J.BIOTECHADV.2021.107732.

[106] Nelson, K. M., Dahlin, J. L., Bisson, J., Graham, J., Pauli, G. F., and Walters, M. A. "The Essential Medicinal Chemistry of Curcumin". In: *Journal of Medicinal Chemistry* 60.5 (2017), 1620–1637. DOI: 10.1021/ACS.JMEDCHEM.6B00975.

[107] *European Parliament. Regulation (EC) No 1333/2008 of the European Parliament and of the Council of 16 December 2008 on food additives*. 2008. URL: http://data.europa.eu/eli/reg/2008/1333/oj.

[108] *Food and Drug Administration. Listing of color additives exempt from certification*. 2001. URL: https://www.ecfr.gov/current/title-21/part-73.

[109] *Grand View Research. Curcumin Market Size, Share & Trends Analysis Report By Application (Pharmaceutical, Food, Cosmetics), By Region (North America, Europe, Asia Pacific, CSA, MEA), And Segment Forecasts, 2020 - 2028*. Tech. rep. Grand View Research Inc., San Francisco, USA, 2018. URL: https://www.grandviewresearch.com/industry-analysis/turmeric-extract-curcumin-market.

[110] Heger, M. "Don't discount all curcumin trial data". In: *Nature* 543.7643 (2017), 40–40. DOI: 10.1038/543040C.

[111] Iturbide, M. et al. "An update of IPCC climate reference regions for subcontinental analysis of climate model data: definition and aggregated datasets". In: *Earth System Science Data* 12.4 (2020), 2959–2970. DOI: 10.5194/ESSD-12-2959-2020.

[112] Jiang, T., Ghosh, R., and Charcosset, C. "Extraction, purification and applications of curcumin from plant materials-A comprehensive review". In: *Trends in Food Science & Technology* 112 (2021), 419–430. DOI: 10.1016/J.TIFS.2021.04.015.

[113] Indira Priyadarsini, K. "The Chemistry of Curcumin: From Extraction to Therapeutic Agent". In: *Molecules* 19 (2014). DOI: 10.3390/molecules191220091.

[114] Rodrigues, J. L., Gomes, D., and Rodrigues, L. R. "A Combinatorial Approach to Optimize the Production of Curcuminoids From Tyrosine in *Escherichia coli*". In: *Frontiers in Bioengineering and Biotechnology* 8 (2020), 59. DOI: 10.3389/fbioe.2020.00059.

[115] Katsuyama, Y., Matsuzawa, M., Funa, N., and Horinouchi, S. "Production of Curcuminoids by *Escherichia coli* Carrying an Artificial Biosynthesis Pathway". In: *Microbiology,* 154.9 (2008), 2620–2628. DOI: 10.1099/mic.0.2008/018721-0.

[116] Rodrigues, J. L., Araújo, R. G., Prather, K. L., Kluskens, L. D., and Rodrigues, L. R. "Production of curcuminoids from tyrosine by a metabolically engineered *Escherichia coli* using caffeic acid as an intermediate". In: *Biotechnology Journal* 10.4 (2015), 599–609. DOI: 10.1002/BIOT.201400637.

[117] Katsuyama, Y., Kita, T., and Horinouchi, S. "Identification and Characterization of Multiple Curcumin Synthases from the Herb *Curcuma longa*". In: *FEBS Letters* 583.17 (2009), 2799–2803. DOI: 10.1016/j.febslet.2009.07.029.

[118] Mishra, S., Wang, Z., Volk, M. J., and Zhao, H. "Design and application of a kinetic model of lipid metabolism in *Saccharomyces cerevisiae*". In: *Metabolic Engineering* 75 (2023), 12–18. DOI: 10.1016/J.YMBEN.2022.11.003.

[119] "A dynamic kinetic model captures cell-free metabolism for improved butanol production". In: *Metabolic Engineering* (2023). DOI: 10.1016/J.YMBEN.2023.01.009.

[120] Poblete-Castro, I., Becker, J., Dohnt, K., Martins dos Santos, V., and Wittmann, C. "Industrial Biotechnology of *Pseudomonas putida* and Related Species". In: *Applied Microbiology and Biotechnology* 93.6 (2012), 2279–2290. DOI: 10.1007/s00253-012-3928-0.

[121] Loeschcke, A. and Thies, S. "Engineering of Natural Product Biosynthesis in *Pseudomonas putida*". In: *Current Opinion in Biotechnology* 65 (2020), 213–224. DOI: 10.1016/j.copbio.2020.03.007.

[122] Nikel, P. I., Chavarría, M., Danchin, A., and de Lorenzo, V. "From Dirt to Industrial Applications: *Pseudomonas putida* as a Synthetic Biology Chassis for Hosting Harsh Biochemical Reactions". In: *Current Opinion in Chemical Biology* 34 (2016), 20–29. DOI: 10.1016/j.cbpa.2016.05.011.

[123] Incha, M. R., Thompson, M. G., Blake-Hedges, J. M., Liu, Y., Pearson, A. N., Schmidt, M., Gin, J. W., Petzold, C. J., Deutschbauer, A. M., and Keasling, J. D. "Leveraging Host Metabolism for Bisdemethoxycurcumin Production in *Pseudomonas putida*". In: *Metabolic Engineering Communications* 10 (2020), e00119. DOI: 10.1016/j.mec.2019.e00119.

[124] Green, R. and Rogers, E. J. "Transformation of Chemically Competent *E. coli*". In: *Methods in Enzymology* 529 (2013), 329–336. DOI: 10.1016/B978-0-12-418687-3.00028-8.

[125] Volke, D. C., Friis, L., Wirth, N. T., Turlin, J., and Nikel, P. I. "Synthetic control of plasmid replication enables target- and self-curing of vectors and expedites genome engineering of *Pseudomonas putida*". In: *Metabolic Engineering Communications* 10 (2020), e00126. DOI: 10.1016/J.MEC.2020.E00126.

[126] Damalas, S. G., Batianis, C., Martin-Pascual, M., Lorenzo, V. de, and Martins dos Santos, V. A. "SEVA 3.1: enabling interoperability of DNA assembly among the SEVA, BioBricks and Type IIS restriction enzyme standards". In: *Microbial Biotechnology* 13.6 (2020), 1793–1806. DOI: 10.1111/1751-7915.13609.

[127] Liebermeister, W., Uhlendorf, J., and Klipp, E. "Modular Rate Laws for Enzymatic Reactions: Thermodynamics, Elasticities and Implementation". In: *Bioinformatics* 26.12 (2010), 1528–1534. DOI: 10.1093/bioinformatics/btq141.

[128] Gläser, L., Kuhl, M., Jovanovic, S., Fritz, M., Vögeli, B., Erb, T. J., Becker, J., and Wittmann, C. "A Common Approach for Absolute Quantification of Short Chain CoA Thioesters in Prokaryotic and Eukaryotic Microbes". In: *Microbial Cell Factories* 19.1 (2020), 160. DOI: 10.1186/s12934-020-01413-1.

[129] Wordofa, G. G., Kristensen, M., Schrübbers, L., McCloskey, D., Forster, J., and Schneider, K. "Quantifying the Metabolome of *Pseudomonas taiwanensis* VLB120: Evaluation of Hot and Cold Combined Quenching/Extraction Approaches". In: *Analytical Chemistry* 89.17 (2017), 8738–8747. DOI: 10.1021/acs.analchem.7b00793.

[130] Bennett, B., Kimball, E., and Gao, M. "Absolute Metabolite Concentrations and Implied Enzyme Active Site Occupancy in *Escherichia coli*." In: *Nature Chemical Biology* 5.8 (2009), 593–599. DOI: 10.1038/nchembio.186.

[131] Noor, E., Haraldsdóttir, H. S., Milo, R., and Fleming, R. M. "Consistent Estimation of Gibbs Energy Using Component Contributions". In: *PLoS Computational Biology* 9.7 (2013), e1003098. DOI: 10.1371/journal.pcbi.1003098.

[132] Katsuyama, Y., Miyazono, K.-i., Tanokura, M., Ohnishi, Y., and Horinouchi, S. "Structural and Biochemical Elucidation of Mechanism for Decarboxylative Condensation of β-Keto Acid by Curcumin Synthase". In: *Journal of Biological Chemistry* 286.8 (2011), 6659–6668. DOI: 10.1074/jbc.M110.196279.

[133] Lubitz, T., Schulz, M., Klipp, E., and Liebermeister, W. "Parameter Balancing in Kinetic Models of Cell Metabolism". In: *The Journal of Physical Chemistry B* 114.49 (2010), 16298–16303. DOI: 10.1021/jp108764b.

[134] Beber, M. E., Gollub, M. G., Mozaffari, D., Shebek, K. M., Flamholz, A. I., Milo, R., and Noor, E. "eQuilibrator 3.0: A Database Solution for Thermodynamic Constant Estimation". In: *Nucleic Acids Research* (2021), gkab1106. DOI: 10.1093/nar/gkab1106.

[135] Stapor, P., Weindl, D., Ballnus, B., Hug, S., Loos, C., Fiedler, A., Krause, S., Hroß, S., Fröhlich, F., and Hasenauer, J. "PESTO: Parameter EStimation TOolbox". In: *Bioinformatics* 34.4 (2018), 705–707. DOI: 10.1093/bioinformatics/btx676.

[136] Fröhlich, F., Kaltenbacher, B., Theis, F. J., and Hasenauer, J. "Scalable Parameter Estimation for Genome-Scale Biochemical Reaction Networks". In: *PLOS Computational Biology* 13.1 (2017), e1005331. DOI: 10.1371/journal.pcbi.1005331.

[137] Fang, Z., Jones, J. A., Zhou, J., and Koffas, M. A. G. "Engineering *Escherichia coli* Co-Cultures for Production of Curcuminoids From Glucose". In: *Biotechnology Journal* 13.5 (2018), 1700576. DOI: 10.1002/biot.201700576.

[138] Wang, S., Zhang, S., Xiao, A., Rasmussen, M., Skidmore, C., and Zhan, J. "Metabolic engineering of *Escherichia coli* for the biosynthesis of various phenylpropanoid derivatives". In: *Metabolic Engineering* 29 (2015), 153–159. DOI: 10.1016/J.YMBEN.2015.03.011.

[139] Hamberger, B. and Hahlbrock, K. "The 4-coumarate:CoA ligase gene family in *Arabidopsis thaliana* comprises one rare, sinapate-activating and three commonly occurring isoenzymes". In: *Proceedings of the National Academy of Sciences* 101.7 (2004), 2209–2214. DOI: 10.1073/PNAS.0307307101.

[140] Wu, J., Chen, W., Zhang, Y., Zhang, X., Jin, J. M., and Tang, S. Y. "Metabolic Engineering for Improved Curcumin Biosynthesis in *Escherichia coli*". In: *Journal of Agricultural and Food Chemistry* 68.39 (2020), 10772–10779. DOI: 10.1021/ACS.JAFC.0C04276.

[141] Couto, M. R., Rodrigues, J. L., and Rodrigues, L. R. "Optimization of fermentation conditions for the production of curcumin by engineered *Escherichia coli*". In: *Journal of The Royal Society Interface* 14.133 (2017). DOI: 10.1098/RSIF.2017.0470.

[142] Katsuyama, Y., Matsuzawa, M., Funa, N., and Horinouchi, S. "In vitro synthesis of curcuminoids by type III polyketide synthase from *Oryza sativa*". In: *The Journal of biological chemistry* 282.52 (2007), 37702–37709. DOI: 10.1074/JBC.M707569200.

[143] Rodrigues, J. L., Couto, M. R., Araújo, R. G., Prather, K. L. J., Kluskens, L., and Rodrigues, L. R. "Hydroxycinnamic Acids and Curcumin Production in Engineered *Escherichia coli* Using Heat Shock Promoters". In: *Biochemical Engineering Journal* 125 (2017), 41–49. DOI: 10.1016/j.bej.2017.05.015.

[144] Wu, Y., Chen, T., Liu, Y., Tian, R., Lv, X., Li, J., Du, G., Chen, J., Ledesma-Amaro, R., and Liu, L. "Design of a Programmable Biosensor-CRISPRi Genetic Circuits for Dynamic and Autonomous Dual-Control of Metabolic Flux in *Bacillus subtilis*". In: *Nucleic Acids Research* 48.2 (2020), 996–1009. DOI: 10.1093/nar/gkz1123.

[145] Rainha, J., Rodrigues, J. L., Faria, C., and Rodrigues, L. R. "Curcumin biosynthesis from ferulic acid by engineered *Saccharomyces cerevisiae*". In: *Biotechnology Journal* 17.3 (2022), 2100400. DOI: 10.1002/BIOT.202100400.

[146] Palmer, C. M., Miller, K. K., Nguyen, A., and Alper, H. S. "Engineering 4-coumaroyl-CoA derived polyketide production in *Yarrowia lipolytica* through a β-oxidation mediated strategy". In: *Metabolic Engineering* 57 (2020), 174–181. DOI: 10.1016/J.YMBEN.2019.11.006.

[147] Lee, S. Y., Kim, H. U., Chae, T. U., Cho, J. S., Kim, J. W., Shin, J. H., Kim, D. I., Ko, Y.-S., Jang, W. D., and Jang, Y.-S. "A comprehensive metabolic map for production of bio-based chemicals". en. In: *Nature Catalysis* 2.1 (2019), 18–33. DOI: 10.1038/s41929-018-0212-4.

[148] Cho, J. S., Kim, G. B., Eun, H., Moon, C. W., and Lee, S. Y. "Designing Microbial Cell Factories for the Production of Chemicals". In: *JACS Au* 2.8 (2022). Publisher: American Chemical Society, 1781–1799. DOI: 10.1021/jacsau.2c00344.

[149] Machado, D. and Herrgård, M. J. "Co-Evolution of Strain Design Methods Based on Flux Balance and Elementary Mode Analysis". In: *Metabolic Engineering Communications* 2 (2015), 85–92. DOI: 10.1016/j.meteno.2015.04.001.

[150] Trinh, C. T., Wlaschin, A., and Srienc, F. "Elementary Mode Analysis: A Useful Metabolic Pathway Analysis Tool for Characterizing Cellular Metabolism". In: *Applied Microbiology and Biotechnology* 81.5 (2009), 813–826. DOI: 10.1007/s00253-008-1770-1.

[151] Kamp, A. von and Klamt, S. "Enumeration of Smallest Intervention Strategies in Genome-Scale Metabolic Networks". In: *PLOS Computational Biology* 10.1 (2014), e1003378. DOI: 10.1371/journal.pcbi.1003378.

[152] Burgard, A. P., Pharkya, P., and Maranas, C. D. "Optknock: A Bilevel Programming Framework for Identifying Gene Knockout Strategies for Microbial Strain Optimization". In: *Biotechnology and Bioengineering* 84.6 (2003), 647–657. DOI: 10.1002/bit.10803.

[153] Tepper, N. and Shlomi, T. "Predicting Metabolic Engineering Knockout Strategies for Chemical Production: Accounting for Competing Pathways". In: *Bioinformatics* 26.4 (2010), 536–543. DOI: 10.1093/bioinformatics/btp704.

[154] Patil, K. R., Rocha, I., Förster, J., and Nielsen, J. "Evolutionary Programming as a Platform for *In Silico* Metabolic Engineering". In: *BMC Bioinformatics* 6.1 (2005), 1–12. DOI: 10.1186/1471-2105-6-308.

[155] Jensen, K., Broeken, V., Hansen, A. S. L., Sonnenschein, N., and Herrgård, M. J. "OptCouple: Joint Simulation of Gene Knockouts, Insertions and Medium Modifications for Prediction of Growth-Coupled Strain Designs". In: *Metabolic Engineering Communications* 8 (2019), e00087. DOI: 10.1016/j.mec.2019.e00087.

[156] Ranganathan, S., Suthers, P. F., and Maranas, C. D. "OptForce: An Optimization Procedure for Identifying All Genetic Manipulations Leading to Targeted Overproductions". In: *PLOS Computational Biology* 6.4 (2010), e1000744. DOI: 10.1371/journal.pcbi.1000744.

[157] Jiang, S., Otero-Muras, I., Banga, J. R., Wang, Y., Kaiser, M., and Krasnogor, N. "OptDesign: Identifying Optimum Design Strategies in Strain Engineering for Biochemical Production". In: *ACS Synthetic Biology* (2022). DOI: 10.1021/ACSSYNBIO.1C00610.

[158] Ravi, S. and Gunawan, R. "ΔFBA—Predicting metabolic flux alterations using genome-scale metabolic models and differential transcriptomic data". In: *PLOS Computational Biology* 17.11 (2021), e1009589. DOI: 10.1371/JOURNAL.PCBI.1009589.

[159] Lo, T. M., Chng, S. H., Teo, W. S., Cho, H. S., and Chang, M. W. "A Two-Layer Gene Circuit for Decoupling Cell Growth from Metabolite Production". In: *Cell Systems* 3.2 (2016), 133–143. DOI: 10.1016/j.cels.2016.07.012.

[160] Raj, K., Venayak, N., and Mahadevan, R. "Novel two-stage processes for optimal chemical production in microbes". In: *Metabolic Engineering* 62 (2020), 186–197. DOI: 10.1016/j.ymben.2020.08.006.

[161] Lewis, N. E., Hixson, K. K., Conrad, T. M., Lerman, J. A., Charusanti, P., Polpitiya, A. D., Adkins, J. N., Schramm, G., Purvine, S. O., Lopez-Ferrer, D., Weitz, K. K., Eils, R., König, R., Smith, R. D., and Palsson, B. Ø. "Omic Data from Evolved *E. coli* Are Consistent with Computed Optimal Growth from Genome-Scale Models". In: *Molecular Systems Biology* 6.1 (2010), 390. DOI: 10.1038/msb.2010.47.

[162] Megchelenbrink, W., Huynen, M., and Marchiori, E. "optGpSampler: An Improved Tool for Uniformly Sampling the Solution-Space of Genome-Scale Metabolic Networks". In: *PLoS ONE* 9.2 (2014). Ed. by S. Rogers, e86587. DOI: 10.1371/journal.pone.0086587.

[163] Ebrahim, A., Lerman, J. A., Palsson, B. O., and Hyduke, D. R. "COBRApy: COnstraints-Based Reconstruction and Analysis for Python". In: *BMC Systems Biology* 7.1 (2013), 74. DOI: 10.1186/1752-0509-7-74.

[164] Cowles, M. K. and Carlin, B. P. "Markov Chain Monte Carlo Convergence Diagnostics: A Comparative Review". In: *Journal of the American Statistical Association* 91.434 (1996), 883–904. DOI: 10.1080/01621459.1996.10476956.

[165] Pham, N., Reijnders, M., Suarez-Diez, M., Nijsse, B., Springer, J., Eggink, G., and Schaap, P. J. "Genome-Scale Metabolic Modeling Underscores the Potential of *Cutaneotrichosporon oleaginosus* ATCC 20509 as a Cell Factory for Biofuel Production". In: *Biotechnology for Biofuels* 14.1 (2021), 2. DOI: 10.1186/s13068-020-01838-1.

[166] 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., and Nielsen, J. "A consensus *S. cerevisiae* metabolic model Yeast8 and its ecosystem for comprehensively probing cellular metabolism". In: *Nature Communications* 10.1 (2019), 3586. DOI: 10.1038/s41467-019-11581-3.

[167] Costenoble, R., Picotti, P., Reiter, L., Stallmach, R., Heinemann, M., Sauer, U., and Aebersold, R. "Comprehensive Quantitative Analysis of Central Carbon and Amino-acid Metabolism in *Saccharomyces cerevisiae* under Multiple Conditions by Targeted Proteomics". In: *Molecular Systems Biology* 7.1 (2011), 464. DOI: 10.1038/msb.2010.122.

[168] Lahtvee, P. J., Sánchez, B. J., Smialowska, A., Kasvandik, S., Elsemman, I. E., Gatto, F., and Nielsen, J. "Absolute Quantification of Protein and mRNA Abundances Demonstrate Variability in Gene-Specific Translation Efficiency in Yeast". In: *Cell Systems* 4.5 (2017), 495–504.e5. DOI: 10.1016/j.cels.2017.03.003.

[169] Lastiri-Pancardo, G., Mercado-Hernández, J. S., Kim, J., Jiménez, J. I., and Utrilla, J. "A quantitative method for proteome reallocation using minimal regulatory interventions". In: *Nature Chemical Biology* 16.9 (2020), 1026–1033. DOI: 10.1038/s41589-020-0593-y.

[170] Elsemman, I. E., Rodriguez Prado, A., Grigaitis, P., Garcia Albornoz, M., Harman, V., Holman, S. W., Heerden, J. van, Bruggeman, F. J., Bisschops, M. M., Sonnenschein, N., Hubbard, S., Beynon, R., Daran-Lapujade, P., Nielsen, J., and Teusink, B. "Whole-cell modeling in yeast predicts compartment-specific proteome constraints that drive metabolic strategies". In: *Nature Communications* 13.1 (2022), 1–12. DOI: 10.1038/s41467-022-28467-6.

[171] Zhang, J., Fang, X., Zhu, X.-L., Li, Y., Xu, H.-P., Zhao, B.-F., Chen, L., and Zhang, X.-D. "Microbial Lipid Production by the Oleaginous Yeast *Cryptococcus curvatus* O3 Grown in Fed-Batch Culture". In: *Biomass and Bioenergy* 35.5 (2011), 1906–1911. DOI: 10.1016/j.biombioe.2011.01.024.

[172] Di Fidio, N., Minonne, F., Antonetti, C., and Raspolli Galletti, A. M. "*Cutaneotrichosporon oleaginosus*: A Versatile Whole-Cell Biocatalyst for the Production of Single-Cell Oil from Agro-Industrial Wastes". In: *Catalysts* 11.11 (2021), 1291. DOI: 10.3390/catal11111291.

[173] Bracharz, F., Beukhout, T., Mehlmer, N., and Brück, T. "Opportunities and Challenges in the Development of *Cutaneotrichosporon oleaginosus* ATCC 20509 as a New Cell Factory for Custom Tailored Microbial Oils". In: *Microbial Cell Factories* 16.1 (2017), 178. DOI: 10.1186/s12934-017-0791-9.

[174] Görner, C., Redai, V., Bracharz, F., Schrepfer, P., Garbe, D., and Brück, T. "Genetic Engineering and Production of Modified Fatty Acids by the Non-Conventional Oleaginous Yeast *Trichosporon oleaginosus* ATCC 20509". In: *Green Chemistry* 18.7 (2016), 2037–2046. DOI: 10.1039/C5GC01767J.

[175] Zhang, H., Zhang, L., Chen, H., Chen, Y. Q., Chen, W., Song, Y., and Ratledge, C. "Enhanced Lipid Accumulation in the Yeast *Yarrowia lipolytica* by Over-Expression of ATP:Citrate Lyase from *Mus musculus*". In: *Journal of Biotechnology* 192 (2014), 78–84. DOI: 10.1016/j.jbiotec.2014.10.004.

[176] Yan, F. X., Dong, G. R., Qiang, S., Niu, Y. J., Hu, C. Y., and Meng, Y. H. "Overexpression of △12, △15-Desaturases for Enhanced Lipids Synthesis in *Yarrowia lipolytica*". In: *Frontiers in Microbiology* 11 (2020). DOI: 10.3389/fmicb.2020.00289.

[177] Madzak, C. "*Yarrowia lipolytica* Strains and Their Biotechnological Applications: How Natural Biodiversity and Metabolic Engineering Could Contribute to Cell Factories Improvement". In: *Journal of Fungi* 7.7 (2021). DOI: 10.3390/jof7070548.

[178] Blazeck, J., Liu, L., Knight, R., and Alper, H. S. "Heterologous production of pentane in the oleaginous yeast *Yarrowia lipolytica*". In: *Journal of Biotechnology* 165.3 (2013), 184–194. DOI: 10.1016/j.jbiotec.2013.04.003.

[179] Kim, M., Park, B. G., Kim, E.-J., Kim, J., and Kim, B.-G. "*In Silico* Identification of Metabolic Engineering Strategies for Improved Lipid Production in *Yarrowia lipolytica* by Genome-Scale Metabolic Modeling". In: *Biotechnology for Biofuels* 12.1 (2019), 1–14. DOI: 10.1186/s13068-019-1518-4.

[180] Gottardi, M., Reifenrath, M., Boles, E., and Tripp, J. "Pathway Engineering for the Production of Heterologous Aromatic Chemicals and Their Derivatives in *Saccharomyces cerevisiae*: Bioconversion from Glucose". In: *FEMS Yeast Research* 17.4 (2017), 35. DOI: 10.1093/femsyr/fox035.

[181] Lian, J., Mishra, S., and Zhao, H. "Recent Advances in Metabolic Engineering of *Saccharomyces cerevisiae*: New Tools and Their Applications". In: *Metabolic Engineering* 50 (2018), 85–108. DOI: 10.1016/J.YMBEN.2018.04.011.

[182] Nigam, P. S. and Luke, J. S. "Food Additives: Production of Microbial Pigments and Their Antioxidant Properties". In: *Current Opinion in Food Science* 7 (2016), 93–100. DOI: 10.1016/J.COFS.2016.02.004.

[183] Rodriguez, A., Strucko, T., Stahlhut, S. G., Kristensen, M., Svenssen, D. K., Forster, J., Nielsen, J., and Borodina, I. "Metabolic Engineering of Yeast for Fermentative Production of Flavonoids". In: *Bioresource Technology* 245 (2017), 1645–1654. DOI: 10.1016/j.biortech.2017.06.043.

[184] Lyu, X., Ng, K. R., Lee, J. L., Mark, R., and Chen, W. N. "Enhancement of Naringenin Biosynthesis from Tyrosine by Metabolic Engineering of *Saccharomyces cerevisiae*". In: *Journal of Agricultural and Food Chemistry* 65.31 (2017), 6638–6646. DOI: 10.1021/acs.jafc.7b02507.

[185] Koendjbiharie, J. G. G., Hon, S., Pabst, M., Hooftman, R., Stevenson, D. M., Cui, J., Amador-Noguez, D., Lynd, L. R., Olson, D. G., and van Kranenburg, R. "The Pentose Phosphate Pathway of Cellulolytic Clostridia Relies on 6-Phosphofructokinase Instead of Transaldolase". In: *Journal of Biological Chemistry* 295.7 (2020), 1867–1878. DOI: 10.1074/jbc.RA119.011239.

[186] Chen, Y., Bao, J., Kim, I. K., Siewers, V., and Nielsen, J. "Coupled Incremental Precursor and Co-Factor Supply Improves 3-Hydroxypropionic Acid Production in *Saccharomyces cerevisiae*". In: *Metabolic Engineering* 22 (2014), 104–109. DOI: 10.1016/j.ymben.2014.01.005.

[187] Li, G., Hu, Y., Jan Zrimec, Luo, H., Wang, H., Zelezniak, A., Ji, B., and Nielsen, J. "Bayesian genome scale modelling identifies thermal determinants of yeast metabolism". In: *Nature Communications* 12.1 (2021), 1–12. DOI: 10.1038/s41467-020-20338-2.

[188] Li, S., Zhang, Q., Wang, J., Liu, Y., Zhao, Y., and Deng, Y. "Recent Progress in Metabolic Engineering of *Saccharomyces cerevisiae* for the Production of Malonyl-CoA Derivatives". In: *Journal of Biotechnology* 325 (2021), 83–90. DOI: 10.1016/j.jbiotec.2020.11.014.

[189] Milke, L., Aschenbrenner, J., Marienhagen, J., and Kallscheuer, N. "Production of Plant-Derived Polyphenols in Microorganisms: Current State and Perspectives". In: *Applied Microbiology and Biotechnology* 102.4 (2018), 1575–1585. DOI: 10.1007/s00253-018-8747-5.

[190] Gold, N. D., Gowen, C. M., Lussier, F.-X., Cautha, S. C., Mahadevan, R., and Marin, V. J. "Metabolic Engineering of a Tyrosine-Overproducing Yeast Using Targeted Metabolomics". In: *Microbial Cell Factories* 14 (2015), 73. DOI: 10.1186/s12934-015-0252-2.

[191] Ferreira, R., Skrekas, C., Hedin, A., Sánchez, B. J., Siewers, V., Nielsen, J., and David, F. "Model-Assisted Fine-Tuning of Central Carbon Metabolism in Yeast through dCas9-Based Regulation". In: *ACS Synthetic Biology* 8.11 (2019), 2457–2463. DOI: 10.1021/acssynbio.9b00258.

[192] Schellenberger, J., Lewis, N. E., and Palsson, B. Ø. "Elimination of Thermodynamically Infeasible Loops in Steady-State Metabolic Models". In: *Biophysical Journal* 100.3 (2011), 544–553. DOI: 10.1016/j.bpj.2010.12.3707.

[193] Desouki, A. A., Jarre, F., Gelius-Dietrich, G., and Lercher, M. J. "CycleFreeFlux: Efficient Removal of Thermodynamically Infeasible Loops from Flux Distributions". In: *Bioinformatics* 31.13 (2015), 2159–2165. DOI: 10.1093/bioinformatics/btv096.

[194] Saa, P. A. and Nielsen, L. K. "ll-ACHRB: A Scalable Algorithm for Sampling the Feasible Solution Space of Metabolic Networks". In: *Bioinformatics* 32.15 (2016), 2330–2337. DOI: 10.1093/bioinformatics/btw132.

[195] Saa, P. A., Zapararte, S., Drovandi, C. C., and Nielsen, L. K. "LooplessFluxSampler: an efficient toolbox for sampling the loopless flux solution space of metabolic models". In: *BMC Bioinformatics 2024 25:1* 25.1 (2024), 1–8. DOI: 10.1186/S12859-023-05616-2.

[196] Yu, T., Liu, Q., Wang, X., Liu, X., Chen, Y., and Nielsen, J. "Metabolic reconfiguration enables synthetic reductive metabolism in yeast". In: *Nature Metabolism* 4.11 (2022). DOI: 10.1038/s42255-022-00654-1.

[197] Noor, E., Eden, E., Milo, R., and Alon, U. "Central Carbon Metabolism as a Minimal Biochemical Walk between Precursors for Biomass and Energy". In: *Molecular Cell* 39.5 (2010), 809–820. DOI: 10.1016/j.molcel.2010.08.031.

[198] Stephanopoulos, G. and Vallino, J. J. "Network Rigidity and Metabolic Engineering in Metabolite Overproduction". In: *Science* 252.1984 (1991), 1675. DOI: 10.1126/science.1904627.

[199] Staffas, L., Gustavsson, M., and McCormick, K. "Strategies and policies for the bioeconomy and bio-based economy: An analysis of official national approaches". In: *Sustainability* 5.6 (2013), 2751–2769. DOI: 10.3390/su5062751.

[200] Nielsen, J. and Keasling, J. D. "Engineering Cellular Metabolism". In: *Cell* 164.6 (2016), 1185–1197. DOI: 10.1016/j.cell.2016.02.004.

[201] Orsi, E., Claassens, N. J., Nikel, P. I., and Lindner, S. N. "Optimizing microbial networks through metabolic bypasses". In: *Biotechnology advances* 60.9 (2022), 108035. DOI: 10.1016/j.biotechadv.2022.108035.

[202] Averesch, N. J. and Krömer, J. O. "Metabolic engineering of the shikimate pathway for production of aromatics and derived compounds-Present and future strain construction strategies". In: *Frontiers in Bioengineering and Biotechnology* 6.3 (2018). DOI: 10.3389/fbioe.2018.00032.

[203] Fujiwara, R., Noda, S., Tanaka, T., and Kondo, A. "Metabolic engineering of *Escherichia coli* for shikimate pathway derivative production from glucose–xylose co-substrate". In: *Nature Communications* 11.1 (2020), 1–7. DOI: 10.1038/s41467-019-14024-1.

[204] Li, Z., Wang, H., Ding, D., Liu, Y., Fang, H., Chang, Z., Chen, T., and Zhang, D. "Metabolic engineering of *Escherichia coli* for production of chemicals derived from the shikimate pathway". In: *Journal of Industrial Microbiology and Biotechnology* 47.6-7 (2020), 525–535. DOI: 10.1007/s10295-020-02288-2.

[205] Braga, A. and Faria, N. "Bioprocess Optimization for the Production of Aromatic Compounds With Metabolically Engineered Hosts: Recent Developments and Future Challenges". In: *Frontiers in Bioengineering and Biotechnology* 8.2 (2020). DOI: 10.3389/fbioe.2020.00096.

[206] Kim, S., Lindner, S. N., Aslan, S., Yishai, O., Wenk, S., Schann, K., and Bar-Even, A. "Growth of *E. coli* on formate and methanol via the reductive glycine pathway". In: *Nature Chemical Biology* 16.5 (2020), 538–545. DOI: 10.1038/s41589-020-0473-5.

[207] Nielsen, J. R., Weusthuis, R. A., and Huang, W. E. "Growth-coupled enzyme engineering through manipulation of redox cofactor regeneration". In: *Biotechnology Advances* 63.9 (2023), 108102. DOI: 10.1016/j.biotechadv.2023.108102.

[208] Orsi, E., Claassens, N. J., Nikel, P. I., and Lindner, S. N. "Growth-coupled selection of synthetic modules to accelerate cell factory development". In: *Nature Communications* 12.1 (2021), 1–5. DOI: 10.1038/s41467-021-25665-6.

[209] Yu, T., Zhou, Y. J., Huang, M., Liu, Q., Pereira, R., David, F., and Nielsen, J. "Reprogramming Yeast Metabolism from Alcoholic Fermentation to Lipogenesis". In: *Cell* 174.6 (2018), 1549–1558. DOI: 10.1016/j.cell.2018.07.013.

[210] Gleizer, S., Ben-Nissan, R., Bar-On, Y. M., Antonovsky, N., Noor, E., Zohar, Y., Jona, G., Krieger, E., Shamshoum, M., Bar-Even, A., and Milo, R. "Conversion of *Escherichia coli* to Generate All Biomass Carbon from CO_2". In: *Cell* 179.6 (2019), 1255–1263. DOI: 10.1016/j.cell.2019.11.009.

[211] Chen, F. Y., Jung, H. W., Tsuei, C. Y., and Liao, J. C. "Converting *Escherichia coli* to a Synthetic Methylotroph Growing Solely on Methanol". In: *Cell* 182.4 (2020), 933–946. DOI: 10.1016/j.cell.2020.07.010.

[212] Iacometti, C., Marx, K., Hönick, M., Biletskaia, V., Schulz-Mirbach, H., Dronsella, B., Satanowski, A., Delmas, V. A., Berger, A., Dubois, I., Bouzon, M., Döring, V., Noor, E., Bar-Even, A., and Lindner, S. N. "Activating Silent Glycolysis Bypasses in *Escherichia coli*". In: *BioDesign Research* 2022 (2022), 1–17. DOI: 10.34133/2022/9859643.

[213] Wirth, N. T., Kozaeva, E., and Nikel, P. I. "Accelerated genome engineering of *Pseudomonas putida* by I-SceI—mediated recombination and CRISPR-Cas9 counterselection". In: *Microbial Biotechnology* 13.1 (2020), 233–249. DOI: 10.1111/1751-7915.13396.

[214] Wang, H. H. and Church, G. M. *Multiplexed genome engineering and genotyping methods: Applications for synthetic biology and metabolic engineering*. 1st ed. Vol. 498. Elsevier Inc., 2011, 409–426. ISBN: 9780123851208. DOI: 10.1016/B978-0-12-385120-8.00018-8.

[215] Chen, S., Zhou, Y., Chen, Y., and Gu, J. "fastp: an ultra-fast all-in-one FASTQ preprocessor". In: *Bioinformatics* 34.17 (2018), i884–i890. DOI: 10.1093/BIOINFORMATICS/BTY560.

[216] Barrick, J. E., Colburn, G., Deatherage, D. E., Traverse, C. C., Strand, M. D., Borges, J. J., Knoester, D. B., Reba, A., and Meyer, A. G. "Identifying structural variation in haploid microbial genomes from short-read resequencing data using breseq". In: *BMC Genomics* 15.1 (2014), 1–17. DOI: 10.1186/1471-2164-15-1039.

[217] Nogales, J., Mueller, J., Gudmundsson, S., Canalejo, F. J., Duque, E., Monk, J., Feist, A. M., Ramos, J. L., Niu, W., and Palsson, B. O. "High-quality genome-scale metabolic modelling of *Pseudomonas putida* highlights its broad metabolic capabilities". In: *Environmental Microbiology* 22.1 (2020), 255–269. DOI: 10.1111/1462-2920.14843.

[218] Nikel, P. I., Kim, J., and Lorenzo, V. de. "Metabolic and regulatory rearrangements underlying glycerol metabolism in *Pseudomonas putida* KT2440". In: *Environmental Microbiology* 16.1 (2014), 239–254. DOI: 10.1111/1462-2920.12224.

[219] Batianis, C., Rosmalen, R. van, Major, M., Ee, C. van, Kasiotakis, A., Weusthuis, R. A., and Martins dos Santos, V. A. "A tunable metabolic valve for precise growth control and increased product formation in *Pseudomonas putida*". In: *Metabolic Engineering* (2022), 118159. DOI: 10.1016/j.ymben.2022.10.002.

[220] Wu, Y., Jameel, A., Xing, X. H., and Zhang, C. "Advanced strategies and tools to facilitate and streamline microbial adaptive laboratory evolution". In: *Trends in Biotechnology* 40.1 (2022), 38–59. DOI: 10.1016/j.tibtech.2021.04.002.

[221] Jha, R. K., Narayanan, N., Pandey, N., Bingen, J. M., Kern, T. L., Johnson, C. W., Strauss, C. E., Beckham, G. T., Hennelly, S. P., and Dale, T. "Sensor-Enabled Alleviation of Product Inhibition in Chorismate Pyruvate-Lyase". In: *ACS Synthetic Biology* 8.4 (2019), 775–786. DOI: 10.1021/acssynbio.8b00465.

[222] Koshla, O., Yushchuk, O., Ostash, I., Dacyuk, Y., Myronovskyi, M., Jäger, G., Süssmuth, R. D., Luzhetskyy, A., Byström, A., Kirsebom, L. A., and Ostash, B. "Gene *miaA* for post-transcriptional modification of tRNAXXA is important for morphological and metabolic differentiation in *Streptomyces*". In: *Molecular Microbiology* 112.1 (2019), 249–265. DOI: 10.1111/mmi.14266.

[223] Olekhnovich, I. and Gussin, G. N. "Effects of mutations in the *Pseudomonas putida miaA* gene: Regulation of the *trpE* and *trpGDC* operons in *P. putida* by attenuation". In: *Journal of Bacteriology* 183.10 (2001), 3256–3260. DOI: 10.1128/JB.183.10.3256-3260.2001.

[224] Yu, J. M., Wang, D., Pierson, L. S., and Pierson, E. A. "Disruption of MiaA provides insights into the regulation of phenazine biosynthesis under suboptimal growth conditions in *Pseudomonas chlororaphis* 30-84". In: *Microbiology* 163.1 (2017), 94–108. DOI: 10.1099/mic.0.000409.

[225] Tian, Z. X., Fargier, E., Mac Aogáin, M., Adams, C., Wang, Y. P., and O'Gara, F. "Transcriptome profiling defines a novel regulon modulated by the LysR-type transcriptional regulator MexT in *Pseudomonas aeruginosa*". In: *Nucleic Acids Research* 37.22 (2009), 7546–7559. DOI: 10.1093/nar/gkp828.

[226] Wehrmann, M., Toussaint, M., Pfannstiel, J., Billard, P., and Klebensberger, J. "The cellular response to lanthanum is substrate specific and reveals a novel route for glycerol metabolism in *Pseudomonas putida* KT2440". In: *mBio* 11.2 (2020), 1–14. DOI: 10.1128/mBio.00516-20.

[227] Lenzen, C., Wynands, B., Otto, M., Bolzenius, J., Mennicken, P., Blank, L. M., and Wierckx, N. "High-yield production of 4-hydroxybenzoate from glucose or glycerol by an engineered *Pseudomonas taiwanensis* VLB120". In: *Frontiers in Bioengineering and Biotechnology* 7.6 (2019), 1–17. DOI: 10.3389/fbioe.2019.00130.

[228] Chaves, J. E., Wilton, R., Gao, Y., Munoz, N. M., Burnet, M. C., Schmitz, Z., Rowan, J., Burdick, L. H., Elmore, J., Guss, A., Close, D., Magnuson, J. K., Burnum-Johnson, K. E., and Michener, J. K. "Evaluation of chromosomal insertion loci in the *Pseudomonas putida* KT2440 genome for predictable biosystems design". In: *Metabolic Engineering Communications* 11.7 (2020), e00139. DOI: 10.1016/j.mec.2020.e00139.

[229] Noda, S., Shirai, T., Mori, Y., Oyama, S., and Kondo, A. "Engineering a synthetic pathway for maleate in *Escherichia coli*". In: *Nature Communications* 8.1 (2017), 1–7. DOI: 10.1038/s41467-017-01233-9.

[230] Zhou, Y., Li, Z., Wang, X., and Zhang, H. "Establishing microbial co-cultures for 3-hydroxybenzoic acid biosynthesis on glycerol". In: *Engineering in Life Sciences* 19.5 (2019), 389–395. DOI: 10.1002/elsc.201800195.

[231] Averesch, N. J. and Rothschild, L. J. "Metabolic engineering of *Bacillus subtilis* for production of para-aminobenzoic acid – unexpected importance of carbon source is an advantage for space application". In: *Microbial Biotechnology* 12.4 (2019), 703–714. DOI: 10.1111/1751-7915.13403.

[232] Noda, S., Mori, Y., Fujiwara, R., Shirai, T., Tanaka, T., and Kondo, A. "Reprogramming *Escherichia coli* pyruvate-forming reaction towards chorismate derivatives production". In: *Metabolic Engineering* 67.3 (2021), 1–10. DOI: 10.1016/j.ymben.2021.05.005.

[233] Wang, J., Zhang, R., Zhang, Y., Yang, Y., Lin, Y., and Yan, Y. "Developing a pyruvate-driven metabolic scenario for growth-coupled microbial production". In: *Metabolic Engineering* 55.7 (2019), 191–200. DOI: 10.1016/j.ymben.2019.07.011.

[234] Li, M., Liu, C., Yang, J., Nian, R., Xian, M., Li, F., and Zhang, H. "Common problems associated with the microbial production of aromatic compounds and corresponding metabolic engineering strategies". In: *Biotechnology Advances* 41.3 (2020), 107548. DOI: 10.1016/j.biotechadv.2020.107548.

[235] Zobel, S., Kuepper, J., Ebert, B., Wierckx, N., and Blank, L. M. "Metabolic response of *Pseudomonas putida* to increased NADH regeneration rates". In: *Engineering in Life Sciences* 17.1 (2017), 47–57. DOI: 10.1002/elsc.201600072.

[236] Asin-Garcia, E., Batianis, C., Li, Y., Fawcett, J. D., Jong, I. de, and Santos, V. A. dos. "Phosphite synthetic auxotrophy as an effective biocontainment strategy for the industrial chassis *Pseudomonas putida*". In: *Microbial Cell Factories* 21.1 (2022), 1–17. DOI: 10.1186/s12934-022-01883-5.

[237] Claassens, N. J., Sánchez-Andrea, I., Sousa, D. Z., and Bar-Even, A. "Towards sustainable feedstocks: A guide to electron donors for microbial carbon fixation". In: *Current Opinion in Biotechnology* 50.ii (2018), 195–205. DOI: 10.1016/j.copbio.2018.01.019.

[238] Shaw, A. J., Lam, F. H., Hamilton, M., Consiglio, A., MacEwen, K., Brevnova, E. E., Greenhagen, E., LaTouf, W. G., South, C. R., Van Dijken, H., and Stephanopoulos, G. "Metabolic engineering of microbial competitive advantage for industrial fermentation processes". In: *Science* 353.6299 (2016), 583–586. DOI: 10.1126/science.aaf6159.

[239] Ling, C. et al. "Muconic acid production from glucose and xylose in *Pseudomonas putida* via evolution and metabolic engineering". In: *Nature communications* 13.1 (2022), 4925. DOI: 10.1038/s41467-022-32296-y.

[240] Elmore, J. R., Dexter, G. N., Salvachúa, D., O'Brien, M., Klingeman, D. M., Gorday, K., Michener, J. K., Peterson, D. J., Beckham, G. T., and Guss, A. M. "Engineered *Pseudomonas putida* simultaneously catabolizes five major components of corn stover lignocellulose: Glucose, xylose, arabinose, p-coumaric acid, and acetic acid". In: *Metabolic Engineering* 62.2 (2020), 62–71. DOI: 10.1016/j.ymben.2020.08.001.

[241] Baldazzi, V., Ropers, D., Gouzé, J. L., Gedeon, T., and Jong, H. de. "Resource allocation accounts for the large variability of rate-yield phenotypes across bacterial strains". In: *eLife* 12 (2023). DOI: 10.7554/ELIFE.79815.

[242] Chen, Y. and Nielsen, J. "Energy metabolism controls phenotypes by protein efficiency and allocation". In: *PNAS* 116.35 (2019), 17592–17597. DOI: 10.1073/pnas.1906569116.

[243] Lorenzo, V. de and Couto, J. "The important versus the exciting: reining contradictions in contemporary biotechnology." In: *Microbial biotechnology* 12.1 (2019), 32–34. DOI: 10.1111/1751-7915.13348.

[244] Kampers, L. F., Asin-Garcia, E., Schaap, P. J., Wagemakers, A., and Martins dos Santos, V. A. "From Innovation to Application: Bridging the Valley of Death in Industrial Biotechnology". In: *Trends in Biotechnology* (2021). DOI: 10.1016/j.tibtech.2021.04.010.

[245] Lewis, N. E., Nagarajan, H., and Palsson, B. O. "Constraining the metabolic genotype–phenotype relationship using a phylogeny of *in silico* methods". In: *Nature Reviews Microbiology* 10 (2012), 291–305. DOI: 10.1038/nrmicro2737.

[246] Lopes, H. and Rocha, I. "Genome-scale modeling of yeast: chronology, applications and critical perspectives". In: *FEMS yeast research* 17.5 (2017). DOI: 10.1093/femsyr/fox050.

[247] Gu, C., Bae Kim, G., Jun Kim, W., Uk Kim, H., and Yup Lee, S. "Current status and applications of genome-scale metabolic models". In: *Genome Biology* (2019). DOI: 10.1186/s13059-019-1730-3.

[248] Choi, K. R., Jang, W. D., Yang, D., Cho, J. S., Park, D., and Lee, S. Y. "Systems Metabolic Engineering Strategies: Integrating Systems and Synthetic Biology with Metabolic Engineering". In: *Trends in Biotechnology* (2019). DOI: 10.1016/j.tibtech.2019.01.003.

[249] Meadows, A. L., Karnik, R., Lam, H., Forestell, S., and Snedecor, B. "Application of dynamic flux balance analysis to an industrial *Escherichia coli* fermentation". In: *Metabolic Engineering* 12.2 (2009), 150–160. DOI: 10.1016/j.ymben.2009.07.006.

[250] Hjersted, J. L., Henson, M. A., and Mahadevan, R. "Genome-scale analysis of *Saccharomyces cerevisiae* metabolism and ethanol production in fed-batch culture". In: *Biotechnology and Bioengineering* 97.5 (2007), 1190–1204. DOI: 10.1002/bit.21332.

[251] Hohenschuh, W., Hector, R. E., Chaplen, F., and Murthy, G. S. "Using high-throughput data and dynamic flux balance modeling techniques to identify points of constraint in xylose utilization in *Saccharomyces cerevisiae*". In: *Systems Microbiology and Biomanufacturing* 1 (2020), 3. DOI: 10.1007/s43393-020-00003-x.

[252] Barford, J. P. "A general model for aerobic yeast growth: Batch growth". In: *Biotechnology and Bioengineering* 35.9 (1990), 907–920. DOI: 10.1002/bit.260350908.

[253] Maris, A. van, Pronk, J., and Dijken, v. P. van. "Patent: Pyruvate producing yeast strain". WO2004099425A2. 2008.

[254] Postma, E., Verduyn, C., Scheffers, W. A., and Van Dijken, J. P. "Enzymic Analysis of the Crabtree Effect in Glucose-Limited Chemostat Cultures of *Saccharomyces cerevisiae*". In: *Applied and Environmental Microbiology* 55.2 (1988), 468–477. DOI: 10.1128/aem.55.2.468-477.1989.

[255] Van Hoek, P., Van Dijken, J. P., and Pronk, J. T. "Effect of Specific Growth Rate on Fermentative Capacity of Baker's Yeast". In: *Applied and Environmental Microbiology* 64.11 (1998), 4226–4233. DOI: 10.1128/aem.64.11.4226-4233.1998.

[256] Rieger, M., Kappeli, O., and Fiechter, A. "The Role of Limited Respiration in the Incomplete Oxidation of Glucose by *Saccharomyces cevevisiae*". In: *Journal of General Microbiology* 129 (1983), 653–661. DOI: 10.1099/00221287-129-3-653.

[257] Van Dijken, J. P. et al. "An interlaboratory comparison of physiological and genetic properties of four *Saccharomyces cerevisiae* strains". In: *Enzyme and Microbial Technology* 26.9-10 (2000), 706–714. DOI: 10.1016/S0141-0229(00)00162-9.

[258] Verduyn, C., Postma, E., Scheffers, W. A., and Dijken, J. P. 1. van. "Physiology of *Saccharomyces cerevisiae* in anaerobic glucose-limited chemostat cultures". In: *Journal of General Microbiology* 136 (1990), 395–403. DOI: 10.1099/00221287-136-3-395.

[259] Canelas, A. B., Ras, C., Pierick, A. ten, Gulik, W. M. van, and Heijnen, J. J. "An *in vivo* data-driven framework for classification and quantification of enzyme kinetics and determination of apparent thermodynamic data". In: *Metabolic Engineering* 13.3 (2011), 294–306. DOI: 10.1016/j.ymben.2011.02.005.

[260] Hanly, T. J., Urello, M., and Henson, M. A. "Dynamic flux balance modeling of *S. cerevisiae* and *E. coli* co-cultures for efficient consumption of glucose/xylose mixtures". In: *Applied Microbiology and Biotechnology* 93.6 (2012), 2529–2541. DOI: 10.1007/s00253-011-3628-1.

[261] Dynesen, J., Smits, H. P., Olsson, L., and Nielsen, J. "Carbon catabolite repression of invertase during batch cultivations of *Saccharomyces cerevisiae*: The role of glucose, fructose, and mannose". In: *Applied Microbiology and Biotechnology* 50.5 (1998), 579–582. DOI: 10.1007/s002530051338.

[262] Frick, O. and Wittmann, C. "Characterization of the metabolic shift between oxidative and fermentative growth in *Saccharomyces cerevisiae* by comparative 13C flux analysis". In: *Microbial Cell Factories* 4.1 (2005), 30. DOI: 10.1186/1475-2859-4-30.

[263] Pereira, R., Nielsen, J., and Rocha, I. "Improving the flux distributions simulated with genome-scale metabolic models of *Saccharomyces cerevisiae*". In: *Metabolic Engineering Communications* 3 (2016), 153–163. DOI: 10.1016/j.meteno.2016.05.002.

[264] Heyland, J., Fu, J., and Blank, L. M. "Correlation between TCA cycle flux and glucose uptake rate during respiro-fermentative growth of *Saccharomyces cerevisiae*". In: *Microbiology* 155 (2009), 3827–3837. DOI: 10.1099/mic.0.030213-0.

[265] Gombert, A. K., Moreira, M., Santos, D., Christensen, B., and Nielsen, J. "Network Identification and Flux Quantification in the Central Metabolism of *Saccharomyces cerevisiae* under Different Conditions of Glucose Repression". In: *Journal of bacteriology* 183.4 (2001), 1441–1451. DOI: 10.1128/JB.183.4.1441-1451.2001.

[266] Maaheimo, H., Fiaux, J., Çakar, Z. P., Bailey, J. E., Sauer, U., and Szyperski, T. "Central carbon metabolism of *Saccharomyces cerevisiae* explored by biosynthetic fractional 13C labeling of common amino acids". In: *European Journal of Biochemistry* 268.8 (2001), 2464–2479. DOI: 10.1046/j.1432-1327.2001.02126.x.

[267] Groot, D. H. de, Lischke, J., Muolo, R., Planqué, R., Bruggeman, F. J., and Teusink, B. "The common message of constraint-based optimization approaches: overflow metabolism is caused by two growth-limiting constraints". In: *Cellular and Molecular Life Sciences* 77.3 (2019), 441–453. DOI: 10.1007/s00018-019-03380-2.

[268] Famili, I., Forster, J., Nielsen, J., and Palsson, B. O. "*Saccharomyces cerevisiae* phenotypes can be predicted by using constraint-based analysis of a genome-scale reconstructed metabolic network". In: *PNAS* 100.23 (2003), 13134–13139. DOI: 10.1073/pnas.2235812100.

[269] Plaza, J. and Bogaerts, P. "FBA-based prediction of biomass and ethanol concentration time profiles in *Saccharomyces cerevisiae* fed-bath cultures". In: *IFAC-PapersOnLine* 52.1 (2019), 82–87. DOI: 10.1016/j.ifacol.2019.06.041.

[270] Zhuang, K., Vemuri, G. N., and Mahadevan, R. "Economics of membrane occupancy and respiro-fermentation". In: *Molecular Systems Biology* 7.500 (2011), 1–9. DOI: 10.1038/msb.2011.34.

[271] Medlock, G. L., Moutinho, T. J., and Papin, J. A. "MeduSA: Software to build and analyze ensembles of genome-scale metabolic network reconstructions". In: *PLoS Computational Biology* 16.4 (2020), 1–11. DOI: 10.1371/journal.pcbi.1007847.

[272] Verduyn, C., Postma, E., Scheffers, W. A., and Van Dijken, J. P. "Effect of benzoic acid on metabolic fluxes in yeasts: A continuous-culture study on the regulation of respiration and alcoholic fermentation". In: *Yeast* 8.7 (1992), 501–517. DOI: 10.1002/yea.320080703.

[273] Ebert, B. E., Kurth, F., Grund, M., Blank, L. M., and Schmid, A. "Response of *Pseudomonas putida* KT2440 to increased NADH and ATP demand". In: *Applied and Environmental Microbiology* 77.18 (2011), 6597–6605. DOI: 10.1128/AEM.05588-11.

[274] Poblete-Castro, I., Escapa, I. F., Jäger, C., Puchalka, J., Chi Lam, C. M., Schomburg, D., Prieto, M. A., and Martins dos Santos, V. A. "The metabolic response of *P. putida* KT2442 producing high levels of polyhydroxyalkanoate under single- and multiple-nutrient-limited growth: Highlights from a multi-level omics approach". In: *Microbial Cell Factories* 11.1 (2012), 1–21. DOI: 10.1186/1475-2859-11-34/TABLES/5.

[275] Garcia-Ochoa, F., Gomez, E., and Santos, V. E. "Fluid dynamic conditions and oxygen availability effects on microbial cultures in STBR: An overview". In: *Biochemical Engineering Journal* 164 (2020), 107803. DOI: 10.1016/j.bej.2020.107803.

[276] Weusthuis, R. A., Lamot, I., Oost, J. van der, and Sanders, J. P. "Microbial production of bulk chemicals: development of anaerobic processes". In: *Trends in Biotechnology* 29.4 (2011), 153–158. DOI: 10.1016/J.TIBTECH.2010.12.007.

[277] Kulakowski, S., Banerjee, D., Scown, C. D., and Mukhopadhyay, A. "Improving microbial bioproduction under low-oxygen conditions". In: *Current Opinion in Biotechnology* 84 (2023), 103016. DOI: 10.1016/J.COPBIO.2023.103016.

[278] Wang, Q. and Nomura, C. T. "Monitoring differences in gene expression levels and polyhydroxyalkanoate (PHA) production in *Pseudomonas putida* KT2440 grown on different carbon sources". In: *Journal of Bioscience and Bioengineering* 110.6 (2010), 653–659. DOI: 10.1016/J.JBIOSC.2010.08.001.

[279] Roca, A., Rodríguez-Herva, J. J., Duque, E., and Ramos, J. L. "Physiological responses of *Pseudomonas putida* to formaldehyde during detoxification". In: *Microbial Biotechnology* 1.2 (2008), 158–169. DOI: 10.1111/J.1751-7915.2007.00014.X.

[280] Miller, C. D., Pettee, B., Zhang, C., Pabst, M., McLean, J. E., and Anderson, A. J. "Copper and cadmium: responses in *Pseudomonas putida* KT2440". In: *Letters in Applied Microbiology* 49.6 (2009), 775–783. DOI: 10.1111/J.1472-765X.2009.02741.X.

[281] Możejko-Ciesielska, J. and Serafim, L. S. "Proteomic Response of *Pseudomonas putida* KT2440 to Dual Carbon-Phosphorus Limitation during mcl-PHAs Synthesis". In: *Biomolecules 2019, Vol. 9, Page 796* 9.12 (2019), 796. DOI: 10.3390/BIOM9120796.

[282] Sasnow, S. S., Wei, H., and Aristilde, L. "Bypasses in intracellular glucose metabolism in iron-limited *Pseudomonas putida*". In: *MicrobiologyOpen* 5.1 (2016), 3–20. DOI: 10.1002/MBO3.287.

[283] Ankenbauer, A., Schäfer, R. A., Viegas, S. C., Pobre, V., Voß, B., Arraiano, C. M., and Takors, R. "*Pseudomonas putida* KT2440 is naturally endowed to withstand industrial-scale stress conditions". In: *Microbial Biotechnology* 13.4 (2020), 1145–1161. DOI: 10.1111/1751-7915.13571.

[284] Duuren, J. B. van, Puchałka, J., Mars, A. E., Bücker, R., Eggink, G., Wittmann, C., and Santos, V. A. dos. "Reconciling *in vivo* and *in silico* key biological parameters of *Pseudomonas putida* KT2440 during growth on glucose under carbon-limited condition". In: *BMC Biotechnology* 13.1 (2013), 1–13. DOI: 10.1186/1472-6750-13-93.

[285] Sohn, S. B., Kim, T. Y., Park, J. M., and Lee, S. Y. "*In silico* genome-scale metabolic analysis of *Pseudomonas putida* KT2440 for polyhydroxyalkanoate synthesis, degradation of aromatics and anaerobic survival". In: *Biotechnology Journal* 5.7 (2010), 739–750. DOI: 10.1002/biot.201000124.

[286] Nikel, P. I. and Lorenzo, V. de. "Engineering an anaerobic metabolic regime in *Pseudomonas putida* KT2440 for the anoxic biodegradation of 1,3-dichloroprop-1-ene". In: *Metabolic Engineering* 15.1 (2013), 98–112. DOI: 10.1016/j.ymben.2012.09.006.

[287] Steen, A., Ütkür, F. Ö., Borrero-de Acuña, J. M., Bunk, B., Roselius, L., Bühler, B., Jahn, D., and Schobert, M. "Construction and characterization of nitrate and nitrite respiring *Pseudomonas putida* KT2440 strains for anoxic biotechnical applications". In: *Journal of Biotechnology* 163.2 (2013), 155–165. DOI: 10.1016/J.JBIOTEC.2012.09.015.

[288] Schmitz, S., Nies, S., Wierckx, N., Blank, L. M., and Rosenbaum, M. A. "Engineering mediator-based electroactivity in the obligate aerobic bacterium *Pseudomonas putida* KT2440". In: *Frontiers in Microbiology* 6.4 (2015), 284. DOI: 10.3389/FMICB.2015.00284.

[289] Kampers, L. F., Van Heck, R. G., Donati, S., Saccenti, E., Volkers, R. J., Schaap, P. J., Suarez-Diez, M., Nikel, P. I., and Martins Dos Santos, V. A. "*In silico*-guided engineering of *Pseudomonas putida* towards growth under micro-oxic conditions". In: *Microbial Cell Factories* 18.1 (2019). DOI: 10.1186/s12934-019-1227-5.

[290] Kampers, L. F., Koehorst, J. J., Heck, R. J. van, Suarez-Diez, M., Stams, A. J., and Schaap, P. J. "A metabolic and physiological design study of *Pseudomonas putida* KT2440 capable of anaerobic respiration". In: *BMC Microbiology* 21.1 (2021), 1–15. DOI: 10.1186/S12866-020-02058-1.

[291] Demling, P., Ankenbauer, A., Klein, B., Noack, S., Tiso, T., Takors, R., and Blank, L. M. "*Pseudomonas putida* KT2440 endures temporary oxygen limitations". In: *Biotechnology and Bioengineering* (2021), bit.27938. DOI: 10.1002/BIT.27938.

[292] Ankenbauer, A., Nitschel, R., Teleki, A., Müller, T., Favilli, L., Blombach, B., and Takors, R. "Micro-aerobic production of isobutanol with engineered *Pseudomonas putida*". In: *Engineering in Life Sciences* (2021), elsc.202000116. DOI: 10.1002/elsc.202000116.

[293] Mutyala, S., Li, S., Khandelwal, H., Kong, D. S., and Kim, J. R. "Citrate Synthase Overexpression of *Pseudomonas putida* Increases Succinate Production from Acetate in Microaerobic Cultivation". In: *ACS Omega* 8.29 (2023), 26231–26242. DOI: 10.1021/ACSOMEGA.3C02520.

[294] Nikel, P. I. and Lorenzo, V. de. "Robustness of *Pseudomonas putida* KT2440 as a host for ethanol biosynthesis". In: *New Biotechnology* 31.6 (2014), 562–571. DOI: 10.1016/J.NBT.2014.02.006.

[295] Fernández-Piñar, R., Cámara, M., Soriano, M. I., Dubern, J. F., Heeb, S., Ramos, J. L., and Espinosa-Urgel, M. "PpoR, an orphan LuxR-family protein of *Pseudomonas putida* KT2440, modulates competitive fitness and surface motility independently of N-acylhomoserine lactones". In: *Environmental Microbiology Reports* 3.1 (2011), 79–85. DOI: 10.1111/J.1758-2229.2010.00190.X.

[296] Bratbak, G. and Dundas, I. "Bacterial dry matter content and biomass estimations". In: *Applied and Environmental Microbiology* 48.4 (1984), 755–757. DOI: 10.1128/AEM.48.4.755-757.1984.

[297] "Analysis of *Pseudomonas putida* growth on non-trivial carbon sources using transcriptomics and genome-scale modelling". In: *Environmental Microbiology Reports* 11.2 (2019), 87–97. DOI: 10.1111/1758-2229.12704.

[298] Bushnell, B. "BBMap: A Fast, Accurate, Splice-Aware Aligner". In: (2014). URL: https://www.osti.gov/biblio/1241166.

[299] Bray, N. L., Pimentel, H., Melsted, P., and Pachter, L. "Near-optimal probabilistic RNA-seq quantification". In: *Nature Biotechnology* 34.5 (2016), 525–527. DOI: 10.1038/nbt.3519.

[300] Sayers, E. W. et al. "Database resources of the national center for biotechnology information". In: *Nucleic acids research* 50.D1 (2022), D20–D26. DOI: 10.1093/NAR/GKAB1112.

[301] "Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2". In: *Genome Biology* 15.12 (2014), 1–21. DOI: 10.1186/S13059-014-0550-8/FIGURES/9.

[302] Nueda, M. J., Tarazona, S., and Conesa, A. "Next maSigPro: updating maSigPro bioconductor package for RNA-seq time series". In: *Bioinformatics* 30.18 (2014), 2598–2602. DOI: 10.1093/BIOINFORMATICS/BTU333.

[303] Nakayasu, E. S., Nicora, C. D., Sims, A. C., Burnum-Johnson, K. E., Kim, Y.-M., Kyle, J. E., Matzke, M. M., Shukla, A. K., Chu, R. K., Schepmoes, A. A., et al. "MPLEx: a robust and universal protocol for single-sample integrative proteomic, metabolomic, and lipidomic analyses". In: *MSystems* 1.3 (2016), e00043–16. DOI: 10.1128/msystems.00043-16.

[304] Gao, Y. et al. "High-Throughput Large-Scale Targeted Proteomics Assays for Quantifying Pathway Proteins in *Pseudomonas putida* KT2440". In: *Frontiers in Bioengineering and Biotechnology* 8 (2020). DOI: 10.3389/fbioe.2020.603488.

[305] Rappsilber, J., Mann, M., and Ishihama, Y. "Protocol for micro-purification, enrichment, pre-fractionation and storage of peptides for proteomics using StageTips". In: *Nature Protocols* 2 (2007), 1896–1906. DOI: 10.1038/nprot.2007.261.

[306] Zecha, J., Satpathy, S., Kanashova, T., Avanessian, S. C., Kane, M. H., Clauser, K. R., Mertins, P., Carr, S. A., and Kuster, B. "TMT Labeling for the Masses: A Robust and Cost-efficient, In-solution Labeling Approach". In: *Molecular & Cellular Proteomics* 18.7 (2019), 1468–1478. DOI: 10.1074/mcp.TIR119.001385.

[307] Feng, Y., Bui, T. P. N., Stams, A. J., Boeren, S., Sánchez-Andrea, I., and Vos, W. M. de. "Comparative genomics and proteomics of *Eubacterium maltosivorans*: functional identification of trimethylamine methyltransferases and bacterial microcompartments in a human intestinal bacterium with a versatile lifestyle". In: *Environmental Microbiology* 24.1 (2022), 517–534. DOI: 10.1111/1462-2920.15886.

[308] Bateman, A. et al. "UniProt: the Universal Protein Knowledgebase in 2023". In: *Nucleic Acids Research* 51.D1 (2023), D523–D531. DOI: 10.1093/NAR/GKAC1052.

[309] Tyanova, S., Temu, T., and Cox, J. "The MaxQuant computational platform for MS-based shotgun proteomics". In: *Nature Protocols* 11.12 (2016), 2301–2319. DOI: 10.1038/nprot.2016.136.

[310] Tyanova, S., Temu, T., Sinitcyn, P., Carlson, A., Hein, M. Y., Geiger, T., Mann, M., and Cox, J. "The Perseus computational platform for comprehensive analysis of (prote)omics data". In: *Nature Methods* 13.9 (2016), 731–740. DOI: 10.1038/nmeth.3901.

[311] Huerta-Cepas, J., Szklarczyk, D., Heller, D., Hernández-Plaza, A., Forslund, S. K., Cook, H., Mende, D. R., Letunic, I., Rattei, T., Jensen, L. J., Von Mering, C., and Bork, P. "eggNOG 5.0: a hierarchical, functionally and phylogenetically annotated orthology resource based on 5090 organisms and 2502 viruses". In: *Nucleic Acids Research* 47.D1 (2019), D309–D314. DOI: 10.1093/NAR/GKY1085.

[312] Maere, S., Heymans, K., and Kuiper, M. "BiNGO: a Cytoscape plugin to assess overrepresentation of Gene Ontology categories in Biological Networks". In: *Bioinformatics* 21.16 (2005), 3448–3449. DOI: 10.1093/BIOINFORMATICS/BTI551.

[313] Martins Dos Santos, V. A., Heim, S., Moore, E. R., Strätz, M., and Timmis, K. N. "Insights into the genomic basis of niche specificity of *Pseudomonas putida* KT2440". In: *Environmental Microbiology* 6.12 (2004), 1264–1286. DOI: 10.1111/J.1462-2920.2004.00734.X.

[314] Conesa, A., Nueda, M. J., Ferrer, A., and Taló, M. "maSigPro: a method to identify significantly differential expression profiles in time-course microarray experiments". In: *Bioinformatics* 22.9 (2006), 1096–1102. DOI: 10.1093/bioinformatics/btl056.

[315] Schwartz, C. J., Djaman, O., Imlay, J. A., and Kiley, P. J. "The cysteine desulfurase, IscS, has a major role in in vivo Fe-S cluster formation in *Escherichia coli*". In: *Proceedings of the National Academy of Sciences of the United States of America* 97.16 (2000), 9009–9014. DOI: 10.1073/PNAS.160261497.

[316] Smith, E. E., Sims, E. H., Spencer, D. H., Kaul, R., and Olson, M. V. "Evidence for Diversifying Selection at the Pyoverdine Locus of *Pseudomonas aeruginosa*". In: *Journal of Bacteriology* 187.6 (2005), 2138. DOI: 10.1128/JB.187.6.2138-2147.2005.

[317] García-Hidalgo, J., Brink, D. P., Ravi, K., Paul, C. J., Lidénb, G., and Gorwa-Grauslund, M. F. "Vanillin Production in pseudomonas: Whole-genome sequencing of *Pseudomonas* sp. strain 9.1 and reannotation of *Pseudomonas putida* CalA as a vanillin reductase". In: *Applied and Environmental Microbiology* 86.6 (2020). DOI: 10.1128/AEM.02442-19.

[318] Linzner, N., Loi, V. V., and Antelmann, H. "The Catalase KatA Contributes to Microaerophilic H_2O_2 Priming to Acquire an Improved Oxidative Stress Resistance in *Staphylococcus aureus*". In: *Antioxidants* 11.9 (2022), 1793. DOI: 10.3390/ANTIOX11091793/S1.

[319] Kurbatov, L., Albrecht, D., Herrmann, H., and Petruschka, L. "Analysis of the proteome of *Pseudomonas putida* KT2440 grown on different sources of carbon and energy". In: *Environmental Microbiology* 8.3 (2006), 466–478. DOI: 10.1111/J.1462-2920.2005.00913.X.

[320] Lenhoff, H. "An inverse relationship of the effects of oxygen and iron on the production of fluorescin and cytochrome c by *Pseudomonas fluorescens*". In: *Nature* 199.4893 (1963), 601–602. DOI: 10.1038/199601a0.

[321] Ponraj, P., Shankar, M., Ilakkiam, D., Rajendhran, J., and Gunasekaran, P. "Influence of periplasmic oxidation of glucose on pyoverdine synthesis in *Pseudomonas putida* S11". In: *Applied Microbiology and Biotechnology* 97.11 (2013), 5027–5041. DOI: 10.1007/S00253-013-4737-9.

[322] Carpenter, C. and Payne, S. M. "Regulation of iron transport systems in *Enterobacteriaceae* in response to oxygen and iron availability". In: *Journal of Inorganic Biochemistry* 133 (2014), 110–117. DOI: 10.1016/J.JINORGBIO.2014.01.007.

[323] Ringel, M. T. and Brüser, T. "The biosynthesis of pyoverdines". In: *Microbial Cell* 5.10 (2018), 424. DOI: 10.15698/MIC2018.10.649.

[324] Venkataramanan, K. P., Min, L., Hou, S., Jones, S. W., Ralston, M. T., Lee, K. H., and Papoutsakis, E. T. "Complex and extensive post-transcriptional regulation revealed by integrative proteomic and transcriptomic analysis of metabolite stress response in *Clostridium acetobutylicum*". In: *Biotechnology for Biofuels* 8.1 (2015), 1–29. DOI: 10.1186/S13068-015-0260-9.

[325] Heim, S., Ferrer, M., Heuer, H., Regenhardt, D., Nimtz, M., and Timmis, K. N. "Proteome reference map of *Pseudomonas putida* strain KT2440 for genome expression profiling: distinct responses of KT2440 and *Pseudomonas aeruginosa* strain PAO1 to iron deprivation and a new form of superoxide dismutase". In: *Environmental Microbiology* 5.12 (2003), 1257–1269. DOI: 10.1111/J.1462-2920.2003.00465.X.

[326] Rintala, E., Toivari, M., Pitkänen, J. P., Wiebe, M. G., Ruohonen, L., and Penttilä, M. "Low oxygen levels as a trigger for enhancement of respiratory metabolism in *Saccharomyces cerevisiae*". In: *BMC Genomics* 10.1 (2009), 461. DOI: 10.1186/1471-2164-10-461.

[327] Rabilloud, T. "Membrane proteins and proteomics: Love is possible, but so difficult". In: *Electrophoresis* 30.S1 (2009), S174–S180. DOI: 10.1002/ELPS.200900050.

[328] Masuda, T., Ito, S., and Ohtsuki, S. "Advances in sample preparation for membrane proteome quantification". In: *Drug Discovery Today: Technologies* 39 (2021), 23–29. DOI: 10.1016/J.DDTEC.2021.06.005.

[329] Oxyr Regulates Kata, P., Hishinuma, A., Yuki, M., Fujimura, M., Fukumori, F., Hishinuma, S., Yuki, M., Fujimura, M., and Fukumori, F. "OxyR regulated the expression of two major catalases, KatA and KatB, along with peroxiredoxin, AhpC in *Pseudomonas putida*". In: *Environmental Microbiology* 8.12 (2006), 2115–2124. DOI: 10.1111/J.1462-2920.2006.01088.X.

[330] Lauhon, C. T. and Kambampati, R. "The *iscS* gene in *Escherichia coli* is required for the biosynthesis of 4-thiouridine, thiamin, and NAD". In: *The Journal of biological chemistry* 275.26 (2000), 20096–20103. DOI: 10.1074/JBC.M002680200.

[331] Babaei, M., Borja Zamfir, G. M., Chen, X., Christensen, B., Kristensen, M., Nielsen, J., and Borodina, I. "Metabolic Engineering of *Saccharomyces cerevisiae* for Rosmarinic Acid Production". In: *ACS Synthetic Biology* 9 (2020). DOI: 10.1021/acssynbio.0c00048.

[332] Jeschek, M., Gerngross, D., and Panke, S. "Combinatorial pathway optimization for streamlined metabolic engineering". In: *Current Opinion in Biotechnology* 47 (2017), 142–151. DOI: 10.1016/J.COPBIO.2017.06.014.

[333] Van Lent, P., Schmitz, J., and Abeel, T. "Simulated Design-Build-Test-Learn Cycles for Consistent Comparison of Machine Learning Methods in Metabolic Engineering". In: *ACS Synthetic Biology* (2023). DOI: 10.1021/acssynbio.3c00186.

[334] Sabzevari, M., Szedmak, S., Penttilä, M., Jouhten, P., and Rousu, J. "Strain design optimization using reinforcement learning". In: *PLOS Computational Biology* 18.6 (2022), e1010177. DOI: 10.1371/JOURNAL.PCBI.1010177.

[335] Carbonell, P., Faulon, J. L., and Breitling, R. "Efficient learning in metabolic pathway designs through optimal assembling". In: *IFAC-PapersOnLine* 52.26 (2019), 7–12. DOI: 10.1016/J.IFACOL.2019.12.228.

[336] Martin-Pascual, M., Moreno-Paz, S., Van Rosmalen, R. P., Dorigo, J., Demaria, F., Van Kranenburg, R., Martins, V. A. P., Santos, D., and Suarez-Diez, M. "Model-guided metabolic engineering of curcuminoid Production in *Pseudomonas putida*". In: *bioRxiv* (2024). DOI: 10.1101/2024.02.08.579459.

[337] Grömping, U. "R Package FrF2 for Creating and Analyzing Fractional Factorial 2-Level Designs". In: *Journal of Statistical Software* 56.1 (2014), 1–56. DOI: 10.18637/JSS.V056.I01.

[338] Robinson, C. J. et al. "Rapid prototyping of microbial production strains for the biomanufacture of potential materials monomers". In: *Metabolic Engineering* 60 (2020), 168–182. DOI: 10.1016/J.YMBEN.2020.04.008.

[339] Casini, A. et al. "A Pressure Test to Make 10 Molecules in 90 Days: External Evaluation of Methods to Engineer Biology". In: *Journal of the American Chemical Society* 140.12 (2018), 4302–4316. DOI: 10.1021/JACS.7B13292.

[340] Liu, Q., Yu, T., Li, X., Chen, Y., Campbell, K., Nielsen, J., and Chen, Y. "Rewiring carbon metabolism in yeast for high level production of aromatic chemicals". In: *Nature Communications* 10.1 (2019), 1–13. DOI: 10.1038/s41467.019.12961.5.

[341] Koopman, F., Beekwilder, J., Crimi, B., Houwelingen, A. van, Hall, R. D., Bosch, D., Maris, A. J. van, Pronk, J. T., and Daran, J. M. "*De novo* production of the flavonoid naringenin in engineered *Saccharomyces cerevisiae*". In: *Microbial Cell Factories* 11.1 (2012), 1–15. DOI: 10.1186/1475.2859.11.155.

[342] Rodriguez, A., Kildegaard, K. R., Li, M., Borodina, I., and Nielsen, J. "Establishment of a yeast platform strain for production of p-coumaric acid through metabolic engineering of aromatic amino acid biosynthesis". In: *Metabolic Engineering* 31 (2015), 181–188. DOI: 10.1016/j.ymben.2015.08.003.

[343] Jendresen, C. B., Stahlhut, S. G., Li, M., Gaspar, P., Siedler, S., Förster, J., Maury, J., Borodina, I., and Nielsen, A. T. "Highly active and specific tyrosine ammonia-lyases from diverse origins enable enhanced production of aromatic compounds in bacteria and *Saccharomyces cerevisiae*". In: *Applied and Environmental Microbiology* 81.13 (2015), 4458–4476. DOI: 10.1128/AEM.00405.15.

[344] Combes, J., Imatoukene, N., Couvreur, J., Godon, B., Brunissen, F., Fojcik, C., Allais, F., and Lopez, M. "Intensification of p-coumaric acid heterologous production using extractive biphasic fermentation". In: *Bioresource Technology* 337.June (2021). DOI: 10.1016/j.biortech.2021.125436.

[345] Moreno-Paz, S., Van Der Hoek, R., Eliana, E., Zwartjens, P., Gosiewska, S., Martins, V. A. P., Santos, D., Schmitz, J., and Suárez-Diez, M. "Machine learning-guided optimization of p-coumaric acid production in yeast". In: *bioRxiv* (2023), 2023.11.27.568789. DOI: 10.1101/2023.11.27.568789.

[346] Verwaal, R., Buiting-Wiessenhaan, N., Dalhuijsen, S., and Roubos, J. A. "CRISPR/Cpf1 enables fast and simple genome editing of *Saccharomyces cerevisiae*". In: *Yeast* 35.2 (2018), 201–211. DOI: 10.1002/YEA.3278.

[347] NCBI. *Saccharomyces cerevisiae CEN.PK113-7D contig163, whole genome shotgun - Nucleotide - NCBI*. URL: https://www.ncbi.nlm.nih.gov/nuccore/AEHG01000256.

[348] Gietz, R. D., Schiestl, R. H., Willems, A. R., and Woods, R. A. "Studies on the transformation of intact yeast cells by the LiAc/SS-DNA/PEG procedure". In: *Yeast* 11.4 (1995), 355–360. DOI: 10.1002/YEA.320110408.

[349] Prins, R. C. and Billerbeck, S. "A buffered media system for yeast batch culture growth". In: *BMC Microbiology* 21.1 (2021), 1–9. DOI: 10.1186/s12866-021-02191-5.

[350] Braus, G. H. "Aromatic amino acid biosynthesis in the yeast *Saccharomyces cerevisiae*: A model system for the regulation of a eukaryotic biosynthetic pathway". In: *Microbiological Reviews* 55.3 (1991), 349–370. DOI: 10.1128/mmbr.55.3.349\-370.1991.

[351] So, K. K., Le, N. M. T., Nguyen, N. L., and Kim, D. H. "Improving expression and assembly of difficult-to-express heterologous proteins in *Saccharomyces cerevisiae* by culturing at a sub-physiological temperature". In: *Microbial cell factories* 22.1 (2023), 55. DOI: 10.1186/S12934-023-02065-7.

[352] Daza, P., Medina, S., Rangel, T., Parra Daza, L. E., Medina, L. S., Rangel, A. E. T., Fernández-Niño, M., Mejía-Manzano, L. A., González-Valdez, J., Reyes, L. H., and Fernando González Barrios, A. "Design and Assembly of a Biofactory for (2S)-Naringenin Production in *Escherichia coli*: Effects of Oxygen Transfer on Yield and Gene Expression". In: *Biomolecules* 13.3 (2023), 565. DOI: 10.3390/BIOM13030565.

[353] Combes, J., Imatoukene, N., Couvreur, J., Godon, B., Fojcik, C., Allais, F., and Lopez, M. "An optimized semi-defined medium for p-coumaric acid production in extractive fermentation". In: *Process Biochemistry* 122 (2022), 357–362. DOI: 10.1016/J.PROCBIO.2022.10.021.

[354] Trantas, E., Panopoulos, N., and Ververidis, F. "Metabolic engineering of the complete pathway leading to heterologous biosynthesis of various flavonoids and stilbenoids in *Saccharomyces cerevisiae*". In: *Metabolic Engineering* 11.6 (2009), 355–366. DOI: 10.1016/J.YMBEN.2009.07.004.

[355] Wang, G., Haringa, C., Noorman, H., Chu, J., and Zhuang, Y. "Developing a Computational Framework To Advance Bioprocess Scale-Up". In: *Trends in Biotechnology* 38.8 (2020), 846–856. DOI: 10.1016/j.tibtech.2020.01.009.

[356] Noorman, H. "An industrial perspective on bioreactor scale-down: What we can learn from combined large-scale bioprocess and model fluid studies". In: *Biotechnology Journal* 6.8 (2011), 934–943. DOI: 10.1002/BIOT.201000406.

[357] Zhang, J., Petersen, S. D., Radivojevic, T., Ramirez, A., Pérez-Manríquez, A., Abeliuk, E., Sánchez, B. J., Costello, Z., Chen, Y., Fero, M. J., Martin, H. G., Nielsen, J., Keasling, J. D., and Jensen, M. K. "Combining mechanistic and machine learning models for predictive engineering and optimization of tryptophan metabolism". In: *Nature Communications* 11.1 (2020), 1–13. DOI: 10.1038/s41467-020-17910-1.

[358] Oyetunde, T., Bao, F. S., Chen, J. W., Martin, H. G., and Tang, Y. J. "Leveraging knowledge engineering and machine learning for microbial bio-manufacturing". In: *Biotechnology Advances* 36.4 (2018), 1308–1315. DOI: 10.1016/j.biotechadv.2018.04.008.

[359] Jervis, A. J. et al. "Machine Learning of Designed Translational Control Allows Predictive Pathway Optimization in *Escherichia coli*". In: *ACS Synthetic Biology* 8.1 (2019), 127–136. DOI: 10.1021/acssynbio.8b00398.

[360] Keun Kang, C., Shin, J., Cha, Y., Sun Kim, M., Sun Choi, M., Kim, T., Park, Y.-K., and Jun Choi, Y. "Machine learning-guided prediction of potential engineering targets for microbial production of lycopene". In: *Bioresource Technology* (2022), 128455. DOI: 10.1016/J.BIORTECH.2022.128455.

[361] Dekker, W. J., Wiersma, S. J., Bouwknegt, J., Mooiman, C., and Pronk, J. T. "Anaerobic growth of *Saccharomyces cerevisiae* CEN.PK113-7D does not depend on synthesis or supplementation of unsaturated fatty acids". In: *FEMS Yeast Research* 19.6 (2019), 60. DOI: 10.1093/FEMSYR/FOZ060.

[362] Young, E. M., Gordon, D. B., and Voigt, C. "Patent: Composability and design of parts for large-scale pathway engineering in yeast". US20170159047A9. 2015.

[363] Roubos, J. A. and VAN Noel, N. M. E. "Patent: A method for achieving improved polypeptide expression". WO2008000632A1. 2007.

[364] Ciurkot, K., Gorochowski, T. E., Roubos, J. A., and Verwaal, R. "Efficient multiplexed gene regulation in *Saccharomyces cerevisiae* using dCas12a". In: *Nucleic Acids Research* 49.13 (2021), 7775–7790. DOI: 10.1093/NAR/GKAB529.

[365] Wouters, B., Miggiels, P., Bezemer, R., Van Der Cruijsen, E. A., Van Leeuwen, E., Gauvin, J., Houben, K., Babu Sai Sankar Gupta, K., Zuijdwijk, P., Harms, A., Carvalho De Souza, A., and Hankemeier, T. "Automated Segmented-Flow Analysis - NMR with a Novel Fluoropolymer Flow Cell for High-Throughput Screening". In: *Analytical Chemistry* 94.44 (2022), 15350–15358. DOI: 10.1021/ACS.ANALCHEM.2C03038.

[366] Lundberg, S. M., Allen, P. G., and Lee, S.-I. "A Unified Approach to Interpreting Model Predictions". In: *Advances in Neural Information Processing Systems* 30 (2017).

[367] Pedregosa, F. et al. "Scikit-learn: Machine Learning in Python". In: *Journal of Machine Learning Research* 12 (2011), 2825–2830.

[368] Gething, M. J. and Davidson, B. E. "Chorismate mutase/prephenate dehydratase from *Escherichia coli* K12. Effects of chemical modification on the enzymic activities and allosteric inhibition". In: *European Journal of Biochemistry* 78.1 (1977), 111–117. DOI: 10.1111/J.1432\-1033.1977.TB11719.X.

[369] Sampathkumar, P. and Morrison, J. F. "Chorismate mutase-prephenate dehydrogenase from *Escherichia coli*. Purification and properties of the bifunctional enzyme". In: *Biochimica et biophysica acta* 702.2 (1982), 204–211. DOI: 10.1016/0167\-4838(82)90504\-0.

[370] Carquet, M., Pompon, D., and Truan, G. "Transcription interference and ORF nature strongly affect promoter strength in a reconstituted metabolic pathway". In: *Frontiers in Bioengineering and Biotechnology* 3.2 (2015), 132035. DOI: 10.3389/FBIOE.2015.00021.

[371] Woodruff, L. B. A., Gorochowski, T. E., Roehner, N., Mikkelsen, T. S., Densmore, D., Gordon, D. B., Nicol, R., and Voigt, C. A. "Registry in a tube: multiplexed pools of retrievable parts for genetic design space exploration". In: *Nucleic Acids Research* 45.3 (2016), 1553–1565. DOI: 10.1093/nar/gkw1226.

[372] Jiang, T., Li, C., and Yan, Y. "Optimization of a p-Coumaric Acid Biosensor System for Versatile Dynamic Performance". In: *ACS Synthetic Biology* 10.1 (2021), 132–144. DOI: 10.1021/ACSSYNBIO.0C00500.

[373] Rizk, M. L. and Liao, J. C. "Ensemble Modeling for Aromatic Production in *Escherichia coli*". In: *PLOS ONE* 4.9 (2009), e6903. DOI: 10.1371/JOURNAL.PONE.0006903.

[374] Mukherjee, M., Blair, R. H., and Wang, Z. Q. "Machine-learning guided elucidation of contribution of individual steps in the mevalonate pathway and construction of a yeast platform strain for terpenoid production". In: *Metabolic Engineering* 74 (2022), 139–149. DOI: 10.1016/J.YMBEN.2022.10.004.

[375] Coussement, P., Bauwens, D., Maertens, J., and De Mey, M. "Direct Combinatorial Pathway Optimization". In: *ACS Synthetic Biology* 6.2 (2017), 224–232. DOI: 10.1021/acssynbio.6b00122.

[376] Basu, S., Kumbier, K., Brown, J. B., and Yu, B. "Iterative random forests to discover predictive and stable high-order interactions". In: *Proceedings of the National Academy of Sciences of the United States of America* 115.8 (2018), 1943–1948. DOI: 10.1073/PNAS.1711236115.

[377] Braniff, N. and Ingalls, B. "New opportunities for optimal design of dynamic experiments in systems and synthetic biology". In: *Current Opinion in Systems Biology* 9 (2018), 42–48. DOI: 10.1016/J.COISB.2018.02.005.

[378] Cuperlovic-Culf, M., Nguyen-Tran, T., and Bennett, S. A. "Machine Learning and Hybrid Methods for Metabolic Pathway Modeling". In: *Methods in Molecular Biology* 2553 (2023), 417–439. DOI: 10.1007/978-1-0716-2617-7_18.

[379] Choi, K., Hellerstein, J., Wiley, H. S., and Sauro, H. M. "Inferring Reaction Networks using Perturbation Data". In: *bioRxiv* (2018), 351767. DOI: 10.1101/351767.

[380] Cao, M., Tran, V. G., and Zhao, H. "Unlocking nature's biosynthetic potential by directed genome evolution". In: *Current Opinion in Biotechnology* 66 (2020), 95–104. DOI: 10.1016/j.copbio.2020.06.012.

[381] Pereira, F., Lopes, H., Maia, P., Meyer, B., Nocon, J., Jouhten, P., Konstantinidis, D., Kafkia, E., Rocha, M., Kötter, P., Rocha, I., and Patil, K. R. "Model-guided development of an evolutionarily stable yeast chassis". In: *Molecular Systems Biology* 17.7 (2021), e10253. DOI: 10.15252/MSB.202110253.

[382] Gernaey, K. V., Lantz, A. E., Tufvesson, P., Woodley, J. M., and Sin, G. "Application of mechanistic models to fermentation and biocatalysis for next-generation processes". In: *Trends in Biotechnology* 28.7 (2010), 346–354. DOI: 10.1016/j.tibtech.2010.03.006.

[383] Chang, L., Liu, X., and Henson, M. A. "Nonlinear model predictive control of fed-batch fermentations using dynamic flux balance models". In: *Journal of Process Control* 42 (2016), 137–149. DOI: 10.1016/J.JPROCONT.2016.04.012.

[384] Lao-Martil, D., Schmitz, J. P., Teusink, B., and Riel, N. A. van. "Elucidating yeast glycolytic dynamics at steady state growth and glucose pulses through kinetic metabolic modeling". In: *Metabolic Engineering* 77 (2023), 128–142. DOI: 10.1016/J.YMBEN.2023.03.005.

[385] Deckwer, W. D., Jahn, D., Hempel, D., and Zeng, A. P. "Systems Biology Approaches to Bioprocess Development". In: *Engineering in Life Sciences* 6.5 (2006), 455–469. DOI: 10.1002/ELSC.200620153.

[386] Weber, P., Kramer, A., Dingler, C., and Radde, N. "Trajectory-oriented Bayesian experiment design versus Fisher A-optimal design: an in depth comparison study". In: *Bioinformatics* 28.18 (2012), i535–i541. DOI: 10.1093/BIOINFORMATICS/BTS377.

[387] Carbonell, P., Radivojevic, T., and García Martín, H. "Opportunities at the Intersection of Synthetic Biology, Machine Learning, and Automation". In: *ACS Synthetic Biology* 8.7 (2019), 1474–1477. DOI: 10.1021/ACSSYNBIO.8B00540.

[388] Baba, T., Ara, T., Hasegawa, M., Takai, Y., Okumura, Y., Baba, M., Datsenko, K. A., Tomita, M., Wanner, B. L., and Mori, H. "Construction of *Escherichia coli* K-12 in-frame, single-gene knockout mutants: The Keio collection". In: *Molecular Systems Biology* 2 (2006). DOI: 10.1038/MSB4100050/SUPPL_FILE/MSB4100050-SUP-0010.XLS.

[389] Liu, Z., Chen, M., Hu, J., Wang, Y., and Chen, Y. "Dynamic minimization of proteome reallocation explains metabolic transition in hierarchical utilization of mixed carbon sources". In: *bioRxiv* (2024), 2024.01.23.576957. DOI: 10.1101/2024.01.23.576957.

[390] Zelezniak, A., Vowinckel, J., Capuano, F., Messner, C. B., Demichev, V., Polowsky, N., Mülleder, M., Kamrad, S., Klaus, B., Keller, M. A., and Ralser, M. "Machine Learning Predicts the Yeast Metabolome from the Quantitative Proteome of Kinase Knockouts". In: *Cell Systems* 7.3 (2018), 269–283.e6. DOI: 10.1016/j.cels.2018.08.001.

[391] Heckmann, D., Lloyd, C. J., Mih, N., Ha, Y., Zielinski, D. C., Haiman, Z. B., Desouki, A. A., Lercher, M. J., and Palsson, B. O. "Machine learning applied to enzyme turnover numbers reveals protein structural correlates and improves metabolic models". In: *Nature Communications* 9.1 (2018), 1–10. DOI: 10.1038/s41467-018-07652-6.

[392] Yu, H., Deng, H., He, J., Keasling, J. D., and Luo, X. "UniKP: a unified framework for the prediction of enzyme kinetic parameters". In: *Nature Communications* 14.1 (2023), 1–13. DOI: 10.1038/s41467-023-44113-1.

[393] Chakrabarti, A., Miskovic, L., Soh, K. C., and Hatzimanikatis, V. "Towards kinetic modeling of genome-scale metabolic networks without sacrificing stoichiometric, thermodynamic and physiological constraints". In: *Biotechnology Journal* 8.9 (2013), 1043–1057. DOI: 10.1002/BIOT.201300091.

[394] Chen, Y. and Nielsen, J. "Mathematical modelling of proteome constraints within metabolism". In: *Current Opinion in Systems Biology* (2021). DOI: 10.1016/j.coisb.2021.03.003.

[395] Ma, J., Yu, M. K., Fong, S., Ono, K., Sage, E., Demchak, B., Sharan, R., and Ideker, T. "Using deep learning to model the hierarchical structure and function of a cell". In: *Nature Methods* 15.4 (2018), 290–298. DOI: 10.1038/nmeth.4627.

[396] Yuan, B., Shen, C., Luna, A., Korkut, A., Marks, D. S., Ingraham, J., and Sander, C. "CellBox: Interpretable Machine Learning for Perturbation Biology with Application to the Design of Cancer Combination Therapy". In: *Cell Systems* 12.2 (2021), 128–140.e4. DOI: 10.1016/j.cels.2020.11.013.

[397] Cambray, G., Guimaraes, J. C., and Arkin, A. P. "Evaluation of 244,000 synthetic sequences reveals design principles to optimize translation in *Escherichia coli*". In: *Nature Biotechnology* 36.10 (2018), 1005–1015. DOI: 10.1038/nbt.4238.

[398] Brown, T. B. et al. "Language Models are Few-Shot Learners". In: *Advances in Neural Information Processing Systems* 33 (2020), 1877–1901. URL: https://commoncrawl.org/the-data/.

[399] Khamwachirapithak, P., Sae-Tang, K., Mhuantong, W., Tanapongpipat, S., Zhao, X. Q., Liu, C. G., Wei, D. Q., Champreda, V., and Runguphan, W. "Optimizing Ethanol Production in *Saccharomyces cerevisiae* at Ambient and Elevated Temperatures through Machine Learning-Guided Combinatorial Promoter Modifications". In: *ACS Synthetic Biology* 12 (2023). DOI: 10.1021/ACSSYNBIO.3C00199.

[400] Nikolados, E. M., Wongprommoon, A., Aodha, O. M., Cambray, G., and Oyarzún, D. A. "Accuracy and data efficiency in deep learning models of protein expression". In: *Nature Communications* 13.1 (2022), 1–12. DOI: 10.1038/s41467-022-34902-5.

[401] Rao, X., Li, D., Su, Z., Nomura, C. T., Chen, S., and Wang, Q. "A smart RBS library and its prediction model for robust and accurate fine-tuning of gene expression in *Bacillus* species". In: *Metabolic Engineering* 81 (2024), 1–9. DOI: 10.1016/J.YMBEN.2023.11.002.

[402] Chory, E. J., Gretton, D. W., DeBenedictis, E. A., and Esvelt, K. M. "Enabling high-throughput biology with flexible open-source automation". In: *Molecular Systems Biology* 17.3 (2021), e9942. DOI: 10.15252/MSB.20209942.

[403] Wilkinson, M. D. et al. "The FAIR Guiding Principles for scientific data management and stewardship". In: *Scientific Data* 3.1 (2016), 160018. DOI: 10.1038/sdata.2016.18.

[404] Martin, H. G. et al. "Perspectives for self-driving labs in synthetic biology". In: *Current Opinion in Biotechnology* 79 (2023), 102881. DOI: 10.1016/J.COPBIO.2022.102881.

[405] Singh, A. H. et al. "An Automated Scientist to Design and Optimize Microbial Strains for the Industrial Production of Small Molecules". In: *bioRxiv* (2023), 1–47. DOI: 10.1101/2023.01.03.521657.

[406] *Financier Worldwide. Amyris files for Chapter 11 protection*. URL: https://www.financierworldwide.com/amyris-files-for-chapter-11-protection.

[407] Endsley, M. R. "Ironies of artificial intelligence". In: *Ergonomics* (2023). DOI: 10.1080/00140139.2023.2243404.

[408] *Global Biofoundry Alliance*. URL: https://www.biofoundries.org/about.

[409] Tobias, M., Ulle, N., and Shoemaker, T. "Fourteen things you need to know about collaborating with data scientists". In: *Nature* (2023). DOI: 10.1038/D41586-023-02291-4.

[410] *Inner Development Goals. IDG Framework*. URL: https://www.innerdevelopmentgoals.org/framework.

[411] *European Comission. Climate change - July 2023 - - Eurobarometer survey*. 2023. URL: https://europa.eu/eurobarometer/surveys/detail/2954.

[412] *World Economic Forum. Global Risks Report 2023*. Tech. rep. 2023. URL: https://www.weforum.org/publications/global-risks-report-2023/digest/.

[413] Carus, M. and Dammer, L. "The Circular Bioeconomy - Concepts, Opportunities, and Limitations". In: *Industrial Biotechnology* 14.2 (2018), 83–91. DOI: 10.1089/IND.2018.29121.MCA.

[414] Richardson, K. et al. "Earth beyond six of nine planetary boundaries". In: *Science advances* 9.37 (2023). DOI: 10.1126/SCIADV.ADH2458.

[415] Cowan, A. E., Klass, S. H., Winegar, P. H., and Keasling, J. D. "Microbial production of fuels, commodity chemicals, and materials from sustainable sources of carbon and energy". In: *Current Opinion in Systems Biology* (2023), 100482. DOI: 10.1016/J.COISB.2023.100482.

[416] Minnaar, L. S., Kruger, F., Fortuin, J., Hoffmeester, L. J., and Haan, R. den. "Engineering *Saccharomyces cerevisiae* for application in integrated bioprocessing biorefineries". In: *Current Opinion in Biotechnology* 85 (2024), 103030. DOI: 10.1016/J.COPBIO.2023.103030.

[417] Bang, J. and Lee, S. Y. "Assimilation of formic acid and CO_2 by engineered *Escherichia coli* equipped with reconstructed one-carbon assimilation pathways". In: *Proceedings of the National Academy of Sciences of the United States of America* 115.40 (2018), E9271–E9279. DOI: 10.1073/PNAS.1810386115.

[418] Tang, H., Wu, L., Guo, S., Cao, W., Ma, W., Wang, X., Shen, J., Wang, M., Zhang, Q., Huang, M., Luo, X., Zeng, J., Keasling, J. D., and Yu, T. "Metabolic engineering of yeast for the production of carbohydrate-derived foods and chemicals from C1–3 molecules". In: *Nature Catalysis* (2023), 1–14. DOI: 10.1038/s41929-023-01063-7.

[419] Liew, F. E. et al. "Carbon-negative production of acetone and isopropanol by gas fermentation at industrial pilot scale". In: *Nature Biotechnology* 40.3 (2022), 335–344. DOI: 10.1038/s41587-021-01195-w.

[420] *European Parliament. 'Green claims' directive Protecting consumers from greenwashing*. Tech. rep. European Parliament, 2023.

APPENDICES

Resumen para el público general

La biotecnología se inspira en la naturaleza para proteger el medio ambiente y mejorar nuestra calidad de vida. Por ejemplo, la biotecnología permite la producción de alimentos mediante fermentación, el descubrimiento de nuevos medicamentos, o el desarrollo de cultivos resistentes a plagas. En concreto, la biotecnología industrial usa microorganismos, llamados factorías celulares, para producir estos compuestos como alternativa a una producción basada en combustibles fósiles. De esta forma, la biotecnología industrial es una de las herramientas que podemos utilizar en la lucha contra el cambio climático. Sin embargo, diseñar estas factorías celulares no es fácil, principalmente porque no sabemos cómo funcionan exactamente. Puedes pensar en el diseño de un coche en el que un equipo de ingenieros decide qué piezas utilizar y sabe la función de cada una de esas piezas. Ahora piensa en una célula. El diseño de la célula no lo hemos hecho nosotros, se basa en miles de años de evolución y la evolución optimiza el crecimiento, cuánto más se propague una célula, mejor. Los biotecnólogos queremos cambiar este "objetivo celular", no nos interesa que nuestra factoría celular crezca mucho, queremos que produzca mucho. Sin embargo, todavía no llegamos a entender por completo cómo funciona una célula, qué pasa cuando intentamos cambiar su objetivo y, para complicar aún más todo, como las condiciones ambientales en las que la cultivamos (lo que se conoce como bio-proceso) afectan su funcionamiento y, por tanto, la producción de nuestro producto.

Una de las herramientas para entender y optimizar las factorías celulares y los bio-procesos son los ciclos de DBTL (Design-Build-Test-Learn, Diseño-Construcción-Experimentación-Aprendizaje). En estos ciclos, diseñamos las células y los experimentos a realizar para obtener la información deseada, construimos las cepas necesarias, las probamos con las condiciones experimentales fijadas y recogemos los datos generados para informar el siguiente ciclo. Para aprovechar toda la información generada en un ciclo de DBTL, necesitamos poder conectar las fases de diseño y aprendizaje. El uso de modelos matemáticos permite la conexión entre estas fases, nos ayuda a recopilar la información generada en un ciclo de DBTL y a utilizarla para diseñar el siguiente ciclo. En esta tesis he probado diferentes modelos con este objetivo. Algunos de estos modelos se basan en lo que sí sabemos sobre el funcionamiento de la célula (modelos mecanísticos). Por ejemplo, gracias al ADN de una célula sabemos qué reacciones bioquímicas pueden suceder

en su interior y esta información se recoge en modelos metabólicos, utilizados en los **Capítulos 3, 4 y 5**. Además, también podemos utilizar ecuaciones diferenciales que incluyen los mecanismos de acción de algunas enzimas (las proteínas que llevan a cabo esas reacciones) en lo que se conoce como modelos cinéticos, utilizados en el **Capítulo 2**. En vez de usar nuestro conocimiento previo, también podemos utilizar modelos que se basan únicamente en los datos experimentales que recogemos durante nuestros experimentos; estos modelos se denominan modelos basados en datos. En los **Capítulos 7 y 8** he usado este tipo de modelos, concretamente basados en el diseño estadístico de experimentos: en función de la información que quiero obtener, este método me permite decidir qué experimentos realizar. Además, en el **Capítulo 9** he utilizado aprendizaje automático ("machine learning") para obtener la máxima información posible de experimentos realizados (más o menos) al azar. Por último, en el **Capítulo 6**, utilizamos lo que se conoce como "técnicas ómicas" que, básicamente, nos permiten tomar una foto de lo que está pasando en la célula en un momento concreto.

Ahora sabes bastante sobre los métodos que he usado pero, ¿qué he conseguido? En el **Capítulo 2** conseguimos mejorar la producción de curcumina (un colorante alimentario natural) en la bacteria *Pseudomonas putida* gracias al uso de modelos cinéticos. Además, aprendemos que controlar la cantidad de cada enzima dentro de la célula es fundamental para mejorar la producción. En los **Capítulos 3 y 4** utilizamos modelos metabólicos. En el **Capítulo 3** creamos una herramienta que nos ayuda a saber cómo modificar nuestra célula para cambiar su "objetivo" del crecimiento a la producción. En el **Capítulo 4** conseguimos cambiar el funcionamiento de *P. putida* para que, cuando crezca, se vea obligada a producir nuestro producto. De esta forma, no tenemos que cambiar su "objetivo".

Hasta ahora solo nos hemos fijado en nuestro microorganismo, pero sabemos que las condiciones ambientales en las que lo cultivamos también afectan su funcionamiento. Esto lo hemos explorado en los **Capítulos 5 y 6**. En el **Capítulo 5** seguimos utilizando modelos metabólicos pero esta vez añadimos una capa más y estudiamos si estos modelos son capaces de predecir cómo el funcionamiento de *Saccharomyces cerevisiae* (la levadura del pan, el vino y la cerveza) cambia según cambiamos la forma de cultivo. A su vez, en el **Capítulo 6**, estudiamos la respuesta de *P. putida* a concentraciones bajas de oxígeno. Igual que nosotros, esta bacteria necesita oxígeno para respirar y gracias al uso de "técnicas ómicas", descubrimos que es capaz de sobrevivir a bajas concentraciones de oxígeno sin grandes cambios en su funcionamiento. Esto es importante, porque cuando cultivamos nuestras factorías celulares a escala industrial, en grandes reactores, es difícil y costoso asegurar que todas las células tengan el oxígeno que necesitan.

En la última parte de la tesis, pasamos a utilizar los modelos basados en datos. En el **Capítulo 7** utilizamos uno de los modelos desarrollados para la producción de curcumina para evaluar diferentes métodos de diseño de experimentos. En el ordenador podemos fácilmente simular cuál es la producción de curcumina de muchas factorías celulares diferentes. Sin embargo, hacer esto en el laboratorio requiere mucho trabajo. Por eso, en este capítulo construimos nuestras cepas *in silico* (es decir, en el ordenador) y descubrimos cuál es la mejor. Después, reducimos el número

de cepas a construir según los diferentes diseños y los evaluamos en base a su capacidad para encontrar a la mejor cepa. Ahora que sabemos qué diseños son los mejores, podemos volver al laboratorio y probarlos de verdad. Esto es lo que hicimos en el **Capítulo 8**, en base al diseño estadístico de experimentos construimos cepas de *S. cerevisiae* para producir ácido cumárico cambiando factores genéticos y ambientales. No solo conseguimos mejorar la producción de este compuesto si no que, además, confirmamos que los factores genéticos y ambientales interaccionan entre ellos. El problema de este método es que no siempre podemos conseguir todas las cepas que queremos en el laboratorio y, si esto pasa, perdemos información. Como alternativa podemos generar cepas de manera (más o menos) aleatoria y utilizar machine learning para analizar los datos obtenidos. Este es el objetivo del **Capítulo 9** en el que volvemos a producir ácido cumárico con *S. cerevisiae* y además aprendemos cómo generar datos de alta calidad para mejorar el entrenamiento de nuestros modelos.

Esta tesis termina con una discusión general (**Capítulo 10**) en la que considero en qué momento, durante el diseño de factorías celulares y bio-procesos, es más conveniente utilizar cada uno de los modelos con los que he trabajado. Además reflexiono sobre la relación entre optimización y aprendizaje durante el desarrollo de factorías celulares. A veces, cuando estamos optimizando nuestra célula, descubrimos cosas sobre su funcionamiento que no sabíamos y, de igual forma, cada vez que mejoramos nuestro conocimiento sobre el funcionamiento celular, hacemos la optimización de los microorganismos más fácil. También discuto cómo, en mi opinión, la combinación de modelos matemáticos y robótica, en lo que se conoce como "biofoundries" puede acelerar la aplicación real de la biotecnología industrial. A pesar del potencial de esta ciencia, esta disertación acaba resaltando la necesidad de estudiar en detalle su sostenibilidad para poder garantizar un futuro verde para todos.

Samenvatting

Biotechnologie gebruikt en vertaalt de kracht van de natuur in toepassingen die bijdragen aan de kwaliteit van het milieu en het leven. Deze toepassingen omvatten onder andere voedselproductie door middel van fermentatie, het ontdekken van nieuwe medicijnen, en het ontwikkelen van ongediertebestendige gewassen. De industriële biotechnologie richt zich op het gebruik van micro-organismes, ook wel celfabrieken genoemd, om biogebaseerde chemicaliën te produceren als alternatief voor de petrochemische industrie wat bijdraagt aan de strijd tegen klimaatverandering. Het ontwerpen van deze celfabrieken is echter een grote uitdaging, omdat we niet volledig begrijpen hoe een cel werkt en of de beoogde productie het functioneren van de cel beïnvloedt. Daarnaast maakt het gebrek aan voorspelbaarheid van celfabrieken het overbruggen van de kloof tussen bioprocesoptimalisatie op laboratoriumschaal en het industrieel opschalen uitdagend. De complexe wisselwerking tussen de talrijke factoren die de prestaties van celfabrieken beïnvloeden, verhindert ons om accurate voorspellingen te maken van hun gedrag. Het is bijvoorbeeld niet duidelijk hoe micro-organismes zich gedragen als genetische factoren en omgevingsfactoren worden verstoord. Productoptimalisatie is daarom vaak een intensief experimenteel proces dat een lange ontwikkelingstermijn doormaakt.

Cycli van Ontwerpen-Bouwen-Testen-Leren (DBTL) vormen een systematische benadering om microbiële stammen en processen te ontwerpen om de productiviteit van een biologische systeem iteratief te verbeteren. Binnen deze cycli worden cellen en experimenten ontworpen om de benodigde informatie te verwerven, stammen te bouwen en te testen onder de gespecificeerde condities, en data te verzamelen en te analyseren om de ontwerpfase van de volgende cyclus te sturen. Om alle informatie die is verkregen gedurende deze cycli te benutten, moeten de ontwerp- en leerfases efficiënt aan elkaar worden gekoppeld. Mathematisch modelleren met zowel kennis- als datagedreven modellen helpen om dit te bereiken. In dit proefschrift worden kennis- en datagedreven modellen zoals kinetische modellen (**Hoofdstuk 2**) en genoomschaal metabole modellen (**Hoofdstukken 3, 4 en 5**) gecombineerd met de analyse van omics data (**Hoofdstuk 6**) en het gebruik van statistisch ontwerp van experimenten (**Hoofdstukken 7 en 8**) en kunstmatige intelligentie (**Hoofdstuk 9**).

Hoofdstukken 2 tot 4 gebruiken mechanistische modellen om de productie van curcumine – een molecuul met toepassingen in de voedingsindustrie en de farmaceutische industrie – te optimaliseren, metabole ontwerpstrategieën te ontwerpen om groei en productie los te koppelen, en

om het bouwen van *Pseudomonas putida* stammen te leiden met een nieuw-voor-de-natuur koolstofkatabolisme waarmee de productie van shikimaat verbeterd wordt. In **Hoofdstuk 2** hebben we de dynamische routemodellering, het systematisch testen van iso-enzymen en de optimalisatie van genexpressieniveaus en substraatconcentraties in *P. putida* gebruikt om de biosynthese van curcuminoïden te verbeteren. Het gebruik van kinetische ensemblemodellen leidde het ontwerp van productiestammen in goede banen, wat de nadruk legde op het belang van het afstemmen van genexpressie. De maximale curcumine-opbrengst van de geoptimaliseerde stam uit tyrosine bedroeg 10,8 $\pm$ 1,8%. Deze opbrengst komt neer op een 4,1-voudige toename in productie-efficiëntie en is de hoogste gerapporteerde opbrengst tot nu toe. **Hoofdstuk 3** introduceert Comparative Flux Sampling Analysis (CFSA), een methode voor het ontwerpen van stammen die volledige metabole ruimtes, geassocieerd met (bijna)-maximale groei- en productiefenotypes met elkaar vergelijkt op basis van genoomschaal metabole modellen. Deze vergelijking, ondersteund door statistische analyses, identificeert reacties met een veranderde flux en draagt doelen aan voor genetische interventies zoals up-regulatie, down-regulatie en gendeleties. CFSA werd toegepast op de productie van lipiden door *Cutaneotrichosporon oleaginosus* en naringenine door *Saccharomyces cerevisiae*. Technische doelstellingen die consistent waren met eerdere studies werden succesvol geïdentificeerd en nieuwe interventies werden voorgesteld. **Hoofdstuk 4** maakt vervolgens gebruik van een combinatie van metabool modelleren, rationeel ontwerp en adaptieve laboratoriumevolutie om het bacteriële metabolisme grondig te herstructureren. Concreet werd in *P. putida* een shikimaatafhankelijke katabole route gecreëerd door de shikimaatroute te herprogrammeren als de dominante route voor groei. Door het sequencen van het hele genoom van de ontwikkelde stammen werden *miaA* en *mexT* geïdentificeerd als sleutelregulatoren. Het verwijderen van deze regulatoren resulteerde in verhoogde fluxen langs de shikimaatroute. De verkregen stam leidt het grootste deel van zijn koolstofkatabolisme langs de shikimaatroute en produceert daarmee 0,35 mol/mol 4-hydroxybenzoaat in minimaal glycerolmedium in de groeifase, waarmee 89,2% van de maximaal voorspelde opbrengst van de route wordt bereikt. Deze hoofdstukken bewijzen het potentieel van kennis gedreven modellen voor de productieoptimalisatie van metabolieten. Terwijl hoofdstukken 2 en 4 de veelzijdigheid van het metabolisme van *P. putida* en het potentieel ervan voor de productie van complexe verbindingen benadrukken, biedt hoofdstuk 3 een robuuste, eenvoudig te gebruiken, gastheer-onafhankelijke methode voor het ontwikkelen van metabole ontwerpstrategieën.

In de **hoofdstukken 5 en 6** verandert de focus naar het begrijpen van de adaptatie van cellen in bioreactoren op het niveau van transcriptoom, proteoom en metaboloom. In **Hoofdstuk 5** gebruikten we flux balans analyse (FBA) en de dynamische FBA (dFBA) om de groeicurve van *S. cereviseae* onder verschillende industriële condities te voorspellen. Deze FBA en dFBA werden toegepast op een genoomschaal model (GEM) en zijn enzyme-gecontroleerde evenknie (ecGEM), waarbij bleek dat de beste voorspelling werd verkregen met een ecGEM. De combinatie van ecGEM met dFBA en het bemonsteren van fluxen faciliteerde het linken van bioreactorcondities en genetische modificatie met voorspellingen van de fluxen. Dit maakte het mogelijk om

de opbrengst en productiviteit voor verschillende stammen en dynamische productieprocessen te voorspellen. Daarnaast suggereerde de toepassing van deze methode dat beperkingen van het proteoom bijdragen aan koolstofcatabolietrepressie. **Hoofdstuk 6** beschrijft de respons van *P. putida* cellen op zuurstof- en glucoselimitaties door middel van chemostaat cultivaties in combinatie met transcriptoom- en proteoomanalyses uitgevoerd. De biomassaopbrengst van langzaam groeiende cellen was 59% hoger onder een zuurstoftekort vergeleken met een glucosetekort als gevolg van de afwezigheid van pyoverdine productie. Deze analyse identificeerde ook 923 genen waarvan het expressieniveau toe- of afgenomen was als gevolg van het zuurstoftekort, terwijl slechts 7 eiwitten op- of neerwaarts gereguleerd waren. Deze observaties duiden op de veerkracht van *P. putida* tijdens lange-termijn zuurstoflimitaties. Zowel hoofdstuk 5 als hoofdstuk 6 benadrukken hoe de groeicondities de celfysiologie beïnvloeden en het belang van de gerelateerde factoren op het ontwerp van de stam en het proces.

In de **hoofdstukken 7 tot en met 9** onderzoek ik het toepassen van data gedreven modellen voor het optimaliseren van celfabrieken en het begrijpen van relaties tussen factoren die de productie van stammen bepalen. In **Hoofdstuk 7** wordt een theoretische studie over het ontwerp van experimenten voor het optimaliseren van metabole routes gepresenteerd. Voortbordurend op een kinetisch model ontwikkeld in hoofdstuk 2 wordt hier de toepasbaarheid van een volledig factoriele ontwerp bibliotheek vergeleken met de resolutie VI, IV, III en Plackett Burman ontwerpen. Onze aanbeveling is om resolutie IV ontwerpen te gebruiken voor het optimaliseren van de expressie van genen die de metabole route beïnvloeden. De ontwerpen maken het mogelijk om optimale stammen te identificeren en geven inzicht wat betreft de impact van factoren en hun interacties op productie. Dit leidt volgende cycli in goede banen. De conclusies die werden getrokken in Hoofdstuk 7 worden toegepast in **Hoofdstuk 8** waar ik DoE gebruikte voor het systematische onderzoeken van de relaties tussen de media, het proces en genetische factoren voor het optimaliseren van de productie van *p*-coumarinezuur. Het laatstgenoemde metaboliet is een uitgangsstof voor meerdere biologische relevante moleculen in *S. cerevisiae*. Het doorlopen van twee cycli van fractionele, factoriele ontwerpen identificeerde factoren die een significant en substantieel effect hadden op de productie van *p*-coumarinezuur. Vervolgens liet de studie een significante interactie zien tussen de temperatuur van de cultuur en de expressie van ARO4. Dit benadrukt het belang van het gelijktijdig optimaliseren van het proces en de stam. Dit bleek ook uit het werk dat is gerapporteerd in de hoofdstukken 5 en 6. In **Hoofdstuk 9** blijven we bij het doel om pCA productie van *S. cerevisiae* te verhogen, maar hier pleiten we voor het genereren van een random bibliotheek van stammen met verschillende genen en expressiegehaltes en hun analyse door middel van kunstmatige intelligentie. Het toepassen van een aanpak die de DNA-sequentie eerste verkende en vervolgens vaststelde, resulteerde in gelaagdheid gedurende het fitten wat de voorspellingen die gebruikmaakten van kleine datasets verbeterde. Vervolgens werden eenvoudige technieken die waren gebaseerd op kunstmatige intelligentie toegepast voor de uitbreiding van de aanvankelijke ruimte van het ontwerp. Deze aanpak resulteerde in een toename van de pCA productie van maar liefst 68% na het doorlopen van twee DBTL cycli.

Deze hoofdstukken onderstrepen de potentie van datagedreven modellen om de ontwerp- en leerfase van de achtereenvolgende DBTL-cycli aan elkaar te relateren, en om het ontwerp van stammen en biotechnische processen te faciliteren.

Deze dissertatie sluit af met een algemene discussie (**Hoofdstuk 10**) die ingaat op de toepassing van de onderzochte modelleeraanpakken voor de verschillende fases gedurende het ontwerp van stammen en biologische processen. Dit betreft met name het aanpassen van metabole routes en het proces. Vervolgens bediscussieer ik de relatie tussen het begrip en procesoptimalisatie in biotechnologisch onderzoek en de impact van geautomatiseerde biofoundries en het modelleren van industriële biotechnologische vraagstukken. Tenslotte reflecteer ik op de behoefte om duurzame substraten te gebruiken en het uitvoeren van levenscyclusanalyses om de duurzaamheid van biologisch geproduceerde producten te evalueren.

Author Affiliations

- Laboratory of Systems and Synthetic Biology - Wageningen University & Research

  Wageningen, The Netherlands.

  Sara Moreno-Paz, María Martín-Pascual, Rik P. van Rosmalen, Julia Dorigo, Francesca Demaria, María Suárez-Diez, Z. Efsun Duman-Özdamar, Lyon Bruinsma, Christos Batianis, Kesi Kurnia, Job J. Dirkmaat, Cristina Furlan, Elif Eliana.

- Laboratory of Microbiology - Wageningen University & Research

  Wageningen, The Netherlands.

  Richard van Kranenburg.

- Bioprocess Engineering Group - Wageningen University & Research,

  Wageningen, The Netherlands

  María Martín-Pascual, Vitor A.P. Martins dos Santos, Z. Efsun Duman-Özdamar, Lyon Bruinsma, Christos Batianis, Ruud A. Weusthuis.

- Lifeglimmer GmbH,

  Berlin, Germany

  Vitor A.P. Martins dos Santos.

- dsm-firmenich,

  Delft, The Netherlands

  Joep Schmitz, Rianne van der Hoek, Priscilla Zwartjens, Silvia Gosiewska.

Completed Training Activities

Discipline Specific	
Python for Data Science - edex	2020
Machine Learning A-Z - Udemy	2020
Machine Learning - BioSB	2020
Integrated modeling and optimization - BioSB	2020
BioSB Conference - BioSB	2020/2021/2023
National Biotechnology Congress - KNCV	2022/2023
Engineering Applications for Microbes - VIB	2022
Metabolic Pathway Analysis - MPA	2023
AI4b.io Annual Symposium - AI4b.io	2024
General	
Competence Assessment - WGS	2020
Focus on Peer Review - Nature	2020
VLAG online lecture series - VLAG	2020
VLAG PhD Week - VLAG	2020
Project Time Management - WGS	2020
Teaching and Supervising Thesis Students - WGS	2021
Career Perspectives - WGS	2022
Biobusiness Summer School - Hyphen Projects	2023
Optional	
Preparation of Research Proposal - VLAG	2020
Group Meetings - SSB	2020-2024
Seminar Series and Journal Clubs - SSB	2020-2024
Assisting in teaching and supervision	
Metabolic Engineering of Industrial Microorganisms	2021-2022
Molecular Systems Biology	2023-2024
iBiosystems	2021
iGEM Supervision	2022
Supervision of BSc and MSc thesis	2020-2023

List of publications

- Moreno-Paz, S., Schmitz, J., Martins dos Santos, V. A., & Suárez-Diez, M. (2022). Enzyme-constrained models predict the dynamics of *Saccharomyces cerevisiae* growth in continuous, batch and fed-batch bioreactors. *Microbial Biotechnology*, *15*(5), 1434-1445.

- Moreno-Paz, S., van der Hoek, R., Eliana, E., Martins dos Santos, V. A., Schmitz, J., & Suárez-Diez, M. (2024). Combinatorial optimization of pathway, process and media for the production of p-coumaric acid by *Saccharomyces cerevisiae*. *Microbial Biotechnology*, *17*(3), e14424.

- Moreno-Paz, S., van der Hoek, R., Eliana, E., Zwartjens, P., Gosiewska, S., Martins dos Santos, V. A., Schmitz, J., & Suárez-Diez, M. (2024). Machine Learning-Guided Optimization of p-Coumaric Acid production in Yeast. *ACS Synthetic Biology*, *13*(*4*), 1312-1322.

- Moreno-Paz, S., Schmitz, J., & Suárez-Diez, M. (2024). *In silico* analysis of Design of Experiment methods for metabolic pathway optimization. *Computational and Structural Biotechnology Journal*. Accepted for publication (2024).

About the author

Sara was born in Plasencia, a town in Cáceres, Spain, on November 15, 1995. She spent her childhood and high school years in the small village of Cuacos de Yuste before she relocated to Salamanca for her Bachelor studies in 2013. There, she pursued a BSc in Biotechnology and graduated *cum laude*. She completed a thesis on the role of xanthine oxidoreductase inhibitors in Chronic Myeloid Leukemia at the Institute for Biomedical Research of Salamanca and an internship focusing on gene regulation in tomatoe at the Institute of Plant Molecular and Cellular Biology in Valencia. Additionally, she spent four months at KU Leuven (Belgium) with an Erasmus scholarship, broadening her perspective on the potential of Biotechnology research and enjoying the international environment. This inspired her to pursue an MSc in Biotechnology with a specialization in Process Technology at Wageningen University in 2017, where she discovered her passion for industrial biotechnology. Her MSc thesis, conducted at the Laboratory of Systems and Synthetic Biology (SSB) and the Bioprocess Engineering Group (BPE), focused on utilizing genome-scale metabolic models and dynamic flux balance analysis for strain design and bioprocess optimization in the production of human milk oligosaccharides. Additionally, she gained practical experience through a six-month internship at Corbion N.V., working on fermentations for the production of a bioplastic monomer. After graduating *cum laude*, she secured funding for her PhD through an NWO Green Top Sector grant in collaboration with María Suárez Diez (SSB), Vitor Martins dos Santos (BPE), and Joep Schmitz (dsm-firmenich). From 2020 to 2024 she worked on this project focusing on modeling strategies for strain and bioprocess design. While Sara's next career move is yet to be determined, she aspires to continue contributing to the field of biotechnology.

Acknowledgments

I cannot believe that the time to write this last section of my thesis is here. This is such a strange feeling that mixes happiness for the achieved and sadness of what's left behind with excitement for the future. Fortunately, happiness and excitement are the predominant feelings and that is mainly due to the lessons learnt during this PhD. Lessons that I did not learn by myself, lessons I learned thanks to all of you.

María, there are no words to describe how supportive, attentive, calm, and smart supervisor you are. I still don't understand how you always find time for all of us and how, although I sometimes felt I was knocking on your door way too often, you always made room for a (not so quick) question. Thank you for caring, well beyond science. I will surely miss our weekly meetings, you are an example to follow and I hope I will once advise people the way you do. I'm looking forward to seeing how SSB will develop with you as chair. **Vitor**, you surely have the ability to see beyond, to verbalize why the science we do matters, I hope I take some of this with me. You are generous and I know you would fight for any of your PhDs. Thank you both because when, back in 2020, I was not feeling strong, I only received support, encouragement, and reassurance from you. When I had to stop, both of you kept reminding me that there was no rush to fully come back. I was writing you made me stronger but, actually, you've made me softer, more flexible, and that is even better. **Joep**, my supervisory team would not have been complete without you. I have enjoyed each of our monthly meetings and looked forward to your always intelligent, helpful, precise feedback that has made all the projects in which we collaborated better. Thank you because, even when strain construction did not always work as expected, you made sure I got the data I needed and helped me reshape our goals. Beyond my three promoters, I would probably not be writing these lines today if it wasn't for **Ruud** and how much I enjoyed working on my MSc thesis. I did not only meet María but also discovered how fun research can be. Thank you all for the advice during these years, you have made me a better scientist.

I would also like to take this opportunity to thank the **members of my thesis committee**. Thank you for taking the time to read and evaluate my work.

One of the aspects of the PhD I enjoyed the most was getting to know me better, understanding how I like to work, and how I can maintain my quality standards without losing myself on the way. Although the end of 2020 was, without a doubt, the hardest part of my PhD, when I look back now, I'm happy I went through it. **Manon**, **Hester**, and **Vesna**, thank you for your support and expertise.

I know I'm lucky to live in a country where mental health matters and support is easy to find as well as to work for an organization that cared and facilitated my recovery.

Another great outcome of my PhD has been meeting my paranymphs: **Sara**, **Efsun**, **Sabine**, **María**, I feel so lucky we were all in SSB at the same time! **Sara**, I remember how, in the middle of COVID and after months of working from home, I asked if you had time for an online coffee break. Since then, everything became so much easier, from the beginning I felt I could trust you, and I started to feel part of SSB. I had someone to look up to, who always has a positive view, and that repeated to me every time I needed that all the stress, and insecurities I felt were normal. Gracias por recordadme que "esto es solo una etapa", tenías razón y espero seguir disfrutando de muchas otras etapas contigo. I don't exactly remember how our online breaks expanded to include **Efsun** and **Sabine**, but I'm so grateful it happened. **Sabine**, I love how you are passionate about so many things with, of course, algae, cats, and cake in the top three. You are always happy and excited and I know I can count on you (even though you prefer cold cheese over cold water). **Efsun**, I feel so identified with you, you represent calm, you make me feel at home, you have a special eye for details and, of course, you are the best host I know. You always have everything under control and it feels safe to be next to you. **María**, there were not many people around when I started going to the lab but I was so lucky you were one of them. You welcomed me, listened to all my concerns, and taught me everything about the lab and much more. No sé cuántas veces me has repetido que "nadie nace sabiendo", tienes una paciencia infinita. María, never forget how amazing, strong, and smart you are, I loved working with you and becoming your friend. You know I'm here for all of you.

Lyon, who cannot be happy when you are around? You are the only person I know who has given serious thoughts on how to create a dragon and I love you for that. Working with you has been amazing and I envy your continuous motivation and unstoppable creativity, even if it includes the creation of couple names. **Marco**, I admire your perseverance, even when things don't work as expected, you are always willing to try again. I'm happy we got to work together. **Christos**, **Enrique**, you were finishing your theses when I started, but somehow I feel you were present during the beginning of my PhD. **Christos**, thank you for creating with me my first putida project (and bearing with it for the following three years) and for making me a part of the SDC team. I admire your confidence and how you are always willing to help, thank you. **Enrique**, I'm in Wageningen partly thanks to you, as back in 2017 you recommended me the Biotech master here. I will always see you as "veterano" and it's a pleasure to witness everything you have achieved.

Once COVID gave as a break, I had the joy to meet the rest of the computational team: **Wasin**, **Rik**, **Sanjee**, I have enjoyed sharing my PhD with you. **Rik**, I always felt I could learn so much from you and I was happy to experience it in the projects we shared. You are generous, accessible, and kind, thank you for taking the time to install so many things on my laptop! **Sanjee**, I'm happy your continuous office moving brought you closer to me, I hope you'll keep me updated on the bright future you have ahead.

Nancy, you were one of the first persons I met in Wageningen, and also the only amateur hairdresser that has ever cut my hair! That shows how much I trust you. It's been great having your office so close to mine and I will miss the lunch break invitations, together with **Maryse**, in my work calendar. Thank you both for being my link to MIB.

Sonja, **Archita**, **Marco**, and all other participants of the PhD trip 2022, it was great getting to know you during our trip to California. **Jenny**, **Claudia**, **Hannah**, **Anne**, **Sam**, **Mike**, **Frank**, and **Iván**, I wish we would have shared more time together but it feels really good to leave SSB feeling the amazing energy you bring, good luck to all of you with your PhDs, I hope I will hear from you!

Nhung, **Erika**, **Henk**, **William**, **Changlin**, and **Bart**, it was a pleasure to share an office with all of you. **Henk**, you have been a constant during my whole PhD and it felt comforting knowing I would find you sitting next to me during these years. Thank you also for the Dutch summary of my thesis! **Peter**, **Edo**, **Jasper**, **Pieter**, **Brett**, **Luis**, and **Alex** thanks for the time we shared in SSB. **Tom**, thank you for your help in the lab, I would not have been able to run reactors without you! **Willemijn**, thank you for making everything easy for all of us, there is no problem you cannot solve. **Cristina**, **Rob**, I think you are doing a great job for PhDs and students to enjoy their time in SSB, I really appreciate your effort. Cristina, thank you for taking care of the lab during difficult times and for your always constructive feedback. I'm happy we could collaborate and even happier to have tried your cooking skills! On the DSM side, I'd like to thank **Priscilla**, **Silvia**, and, especially, **Rianne** for making the projects we had together possible. Rianne, it was great working with you, thank you for always finding the time to answer all my questions. I love the chapters we made together!

Jasper, Foppe, Nanke, Elif, Ally, René, and Emma, I got valuable lessons working with all of you. **Jasper**, it was not easy to work during the pandemic but I'm happy we managed to finish your project together. **Foppe**, from the start, you were already better than me in the lab, it was so easy to work with you. I enjoyed seeing you achieve all your goals, whoever works with you will be extremely lucky. **Nanke**, my little bachelor (of course taller than me), I loved seeing you become more and more independent and how much you learned in so little time. **Elif**, all the work you did during your internship in DSM was impressive and extremely useful and I really enjoyed exchanging research ideas with you. **Ally**, thanks to you I was part of iGEM and I loved seeing how your creative mind became a bit more structured without losing its characteristic sparkle. **René**, you were really fun to work with (stickers in the freezer included), it was great seeing you grow. **Emma**, you were my last student and it was a pleasure seeing your confidence increase in the time we shared together. Thank you all for your contributions to my PhD.

I would like to thank everyone responsible for how much I love living in The Netherlands. **Pedro**, I feel I could spend hours talking to you and never get bored, me transmites paz. Además, contigo aprendo una faceta del español que desconozco. Los pájaros de Cuacos te esperan. **Bea**, we both had a tough start to our PhDs, and I'm so proud of how we turned it around! I loved getting to know you better over the last years. I'm already sad knowing that you are leaving, but I have no doubts we will be reunited in the Peninsula. **Flavia**, I love you, your chaos, and your countless hobbies, although you spend more and more time in Italy, The Netherlands would not be the same

without you. I promise I'll never talk to you about insurance and taxes. **Laura**, fuiste un regalo caído del cielo. Eres la persona más generosa que conozco, con quien me resulta fácil ser vulnerable, a quien no me cuesta pedir consejo. Me encantan nuestras sesiones de catch up y me encanta que nuestra familia española en Utrecht haya crecido. Gracias por presentarme a **Leire** y a **Marina**, ¡es tan fácil reír con vosotras! ¡Os quiero! Y dentro de esa familia está **Silvia**. No sabía en qué parte de los agradecimientos meterte, ¡encajas en todas! Por orden cronológico, gracias por ser una parte central en Salamanca, por, aunque a veces cueste, nunca tomar partidos. Gracias por todas las visitas y, por unos años, convertir Cuacos en tu destino de vacaciones. Y, por supuesto, gracias por ser valiente y venir a Holanda, eres mucho más fuerte de lo que crees y me alegro de tenerte cerca. Gracias también por todos los shake flask experiments con mil strains de curcumina! **Timmy**, **Venda** one of the best things of these years has been getting closer to you, even if it meant studying the names of all the LST students from 2017 to 2019. You've made me love holding hands before a meal, being with you ensures a fun time! Our trip with **Ricardo** and **Melina** was the best way to celebrate the submission of my thesis. **Ricardo**, you are always welcome home and, you know, the more viruses the better. To all of you, and of course, **Nina**, thank you for opening this amazing group to me.

Also, I am so grateful for all of you that, although living further, are always close to me. **Elena**, siempre estás y no me imagino un invierno sin un café en tu salita ni un verano sin conversaciones al lado de la piscina. Porque no importa cuánto tiempo pase pero verte es volver a casa y porque cuando dudo de mi misma siempre estás para recordarme que sí puedo. Gracias, porque tu padre es mi proveedor oficial de pimentón (gracias **Luis**, gracias **Eva**, y, por supuesto, **Helen** and **Robert**, bedankt!). Gracias **Lucía**, **Elena**, **Elvira**, porque aunque a veces cueste cuadrar un día para vernos, cuando lo conseguimos parece que el tiempo no ha pasado. Gracias chichos, por hacer de Salamanca un lugar al que siempre volver y convertir España y Europa en lugares que visitar. **Claudia**, porque da igual coger un avión que comer palomitas en un parque en Béjar; **Ana**, porque recorres los km que haga falta para vernos; **Anabel** por descubir que nos parecemos tanto (e **Irene**, igual que Álvaro); **Dani** porque tu locura me siga haciendo reir. Tenéis un trocito muy grande de mi corazón. Gracias **Diego**, porque volar a Madrid significa verte, por tu alegría, por tu cariño y por recordarme que escribir una tesis puede ser algo muy bonito. Gracias **Mauricio**, por ser un hermano mayor y porque cada vez que pienso que hace tiempo que no hablamos recibo uno de tus audios que adoro. Thank you **Ioanna**, **Jalees**, **Junika**, **Alicia**, **Luca**, for the great time we had and for keep checking on me during these years. I love you all!

Gracias a mi familia, sois lo mejor que tengo. A mis **abuelos**, en especial a abuela **Pepa** y el ejemplo que es para todos nosotros, ojalá parecerme un poco, ojalá un día mi casa sea un lugar de tantos encuentros, buenos recuerdos, y buena comida como es la tuya. Gracias a mis titas y mis tíos: mi **padrino**, **M. Eugenia**, **Graci**, **Celia**, **Uli**, **Mari**, **Nando**, **Rosa**, mi **madrina**, el tío **Eno**, la **tía pequeña**, **Alberto**, **Jose** y **Ana**. Porque soy vuestra sobrina europea, y, cada vez que os veo siento lo orgullosos que estáis de mí. Gracias a **tita Candi**, porque no imagino una mejor representante de abuela en mi defensa y por todas las croquetas; y a **tita Rosi** porque los

tulipanes holandeses no se comparan a los de tu huerta. A mis primos, **Ana**, **Jaime**, **Javier**, **Vera**, **Alberto**, **Patricia**, **Alejandro**, **David**, **Germán** y a mis primitos **Maya**, **Adriana**, **Marco**, y **Lola** porque es una suerte sentiros cerca. Mención especial para Germi, por leerse la intro de mi tesis y Maya, porque no todo el mundo tiene la suerte de encontrar regalos en el buzón. Me encanta hablar con cada uno de vosotros y, aunque me cuesta no estar cuando os reunís todos, se que me teneís tan presente como yo a vosotros. ¡Os quiero!

Mamá, papá, Elvira, todo lo que consiga es gracias a vosotros. Espero que siempre encontremos el tiempo para el FaceTime del finde y, que algún día, las llamadas vuelvan a convertise con más frecuencia en comidas a vuestro lado. **Mamá**, **papá** sois un apoyo incondicional y, aunque a veces no lo parezca, vuestros consejos son siempre los más importantes. Gracias porque siempre nos lo habéis dado todo, educándonos a la vez para saber apreciarlo. Gracias por apoyarme en cada decisión y por la confianza que tenéis en mí. Y gracias porque no os da pereza coger un avión y pasar frío si eso significa pasar unos días juntos. Os admiro a vosotros y todo lo que construís y, sobretodo, os quiero. **Elvira**, para ti unas palabras que me dedicaste a mi: "siempre eliges montañas altas y difíciles, pero también siempre terminas coronándolas y disfrutando de tremendas vistas". Me cuesta creer lo distintas e iguales que somos al mismo tiempo. Tenerte a mi lado es una suerte inmensa. Gracias **Chechu**, por ser parte de la familia, por cuidar y dejarte cuidar por mi hermana y, también, por tus habilidades en reparación de ordenadores, ¡ya hemos comprobado que pueden venir muy bien! Gracias también a **Mariluz** y a **Jose** porque ahora cuando regreso a España tengo otro hogar al que volver.

Álvaro, las gracias más grandes, en mayúsculas, son para ti. Imagino que hacer el doctorado sin ti a mi lado hubiera sido posible pero desde luego mucho más duro, y también mucho más aburrido. Porque he perdido la cuenta de cuántas veces me has levantado, cuántas veces me has animado y cuántas veces me has celebrado. Eso sin contar todas las discusiones científicas que hemos tenido, y todo el feedback que me has dado desde cómo hacer un setup para un chemostat, a mejorar mis presentaciones o ayudarme con el diseño de la portada. Eres una de las personas más inteligentes que conozco y me encanta pensar contigo. Gracias por la suerte que tengo de tenerte a mi lado, porque no hay día que no me hagas reír, y por todos los "todo saldrá bien", somos el mejor equipo. No imagino un compañero mejor, me encanta vernos crecer juntos y, ya lo sabes, un trocito de este PhD es tuyo.

About the Cover

The cover of this thesis symbolizes the parallel scientific and personal development achieved during my PhD. The central motif presents a cell factory (yeast) on a journey from its natural state through a path of successive and continuous improvements until it can leap forward and transcend the horizons as an industrial powerhouse. Escorting it we find a fresh PhD candidate nurturing a plant - herself. The personal growth enabled by the constant and committed care blooms into the brave, yet vulnerable graduate and results in the thesis you are holding.
P.S. Personal growth and scientific progress are seldom achieved alone. The snail represents the patience and support of everyone involved in my PhD journey.

Sobre la portada

El diseño de la portada de esta tesis simboliza el crecimiento logrado durante mi doctorado, tanto científica como personalmente. El concepto central presenta una senda recorrida por una factoría celular (levadura). Comenzando desde su estado natural, sigue un camino de constante y sucesivo progreso hasta poder atravesar el horizonte gracias a su potencial industrial. Acompañándola, se encuentra una joven estudiante cuidando de una planta -ella misma. El desarrollo personal, cimentado por los cuidados constantes, florece dando lugar a una graduada, vulnerable pero valiente, y a la tesis que tienes en tus manos.
P. S. El progreso científico y la realización personal son logros colectivos. El caracol representa el apoyo y la paciencia de todas las personas involucradas en mi doctorado.

The research described in this thesis was financially supported by the Netherlands Organization for Scientific Research (NWO) with project number GSGT.2019.008.

Financial support from Wageningen University for printing this thesis is gratefully acknowledged.

Cover design by Lucía Antruejo
Printed by ProefschriftMaken || www.proefschriftmaken.nl