

Doctoral theses at NTNU, 2021:96

Snorre Sulheim

Assembly and application of
genome-scale metabolic models
to study *Streptomyces coelicolor*
and *Prochlorococcus*

ISBN 9978-82-326-5790-2 (printed ver.)
ISBN 9978-82-326-6517-4 (electronic ver.)
ISSN 1503-8181 (printed ver.)
ISSN 2703-8084 (electronic ver.)

Doctoral theses at NTNU, 2021:96

NTNU
Norwegian University of
Science and Technology
Thesis for the degree of
Philosophiae Doctor
Faculty of Natural Sciences
Department of Biotechnology and Food Science

Snorre Sulheim

**Assembly and application of
genome-scale metabolic models
to study *Streptomyces coelicolor*
and *Prochlorococcus***

Thesis for the degree of Philosophiae Doctor

Trondheim, March 2021

Norwegian University of Science and Technology
Faculty of Natural Sciences
Department of Biotechnology and Food Science



Norwegian University of
Science and Technology

NTNU

Norwegian University of Science and Technology

Thesis for the degree of Philosophiae Doctor

Faculty of Natural Sciences

Department of Biotechnology and Food Science

© Snorre Sulheim

ISBN 978-82-326-5790-2 (printed ver.)

ISBN 978-82-326-6517-4 (electronic ver.)

ISSN 1503-8181 (printed ver.)

ISSN 2703-8084 (electronic ver.)

Doctoral theses at NTNU, 2021:96



Printed by Skipnes Kommunikasjon AS

Abstract

Metabolism is the set of all chemical reactions responsible for the conversion of nutrients into the energy and cellular building blocks required for growth and cellular maintenance in a living organism. Because of our detailed knowledge of enzymes and the chemical reactions they catalyze, one can create a rather accurate representation of an organism's metabolic network from the sequenced and annotated genome. However, in contrast to classical textbook depictions of individual metabolic pathways, these metabolic networks are often highly interconnected and can contain thousands of different reactions and metabolites. Due to this complexity, computational and mathematical algorithms are often required to predict the phenotypic outcome of genetic modifications or changes in the nutrient environment.

When a metabolic network is combined with a representation of growth, cellular maintenance requirements, and available nutrients, it is called a genome-scale metabolic model. In this work we assemble and apply genome-scale metabolic models to study two rather different organisms. The first organism, *Streptomyces coelicolor*, is a complex, soil-dwelling bacterium that is of great interest within drug discovery as a cell factory for production of novel biopharmaceuticals. Through two consecutive publications we merge and improve existing *S. coelicolor* models into a consensus model that is hosted in an open-source environment to encourage contributions from the *Streptomyces* research community. We then apply the developed model to explore and understand how one should proceed with strain development to create a mutant strain that is optimal for heterologous expression of biosynthetic gene clusters. Another contribution in this direction is our development of a computational pipeline that automatically reconstructs metabolic pathways encoded by biosynthetic gene clusters.

The second organism, *Prochlorococcus*, is the most abundant phototrophic marine bacterium, and thus a major player in the marine food web and global carbon fixation. We use random sampling and dynamic flux balance analysis to understand how its metabolism is affected by the day-night cycle and varying nutrient conditions, with a particular focus on glycogen allocation and release of organic compounds that become nutrients for marine heterotrophs. Furthermore, this study required method development extending the software COMETS to account for the periodicity of available daylight and light absorption.

Together, this work contributes to an increased understanding of *S. coelicolor* and *Prochlorococcus*, in addition to updated and improved genome-scale metabolic models which are by themselves valuable tools in further research of these bacteria. Additionally, we have developed generic tools of great value for a broader audience, both towards drug development and for future studies of photosynthetic microbes.

Acknowledgement

This doctoral thesis is submitted to the Norwegian University of Science and Technology (NTNU) as one of the prerequisites for the degree of *philosophiae doctor* (PhD). The work has been conducted over the past four years, both at SINTEF Industry, Department of Biotechnology and Nanomedicine, and at NTNU, Department of Biotechnology and Food Science. Additionally, I had the pleasure of spending six months at Prof. Daniel Segrè's lab in the Bioinformatics Program and Biological Design Center at Boston University, MA, USA. This research stay was partly funded by a travel grant from the Norwegian graduate research school in bioinformatics, biostatistics and systems biology (NORBIS). My PhD position was otherwise funded by SINTEF Industry and the INBioPharm project of the Centre for Digital Life Norway (Research Council of Norway grant no. 248885).

The decision to do this PhD was made in just a few days in the summer of 2016 after a short meeting with Trond Ellingsen and Håvard Sletta. Their immediate trust in my abilities, despite no prior knowledge in biology, was extremely encouraging. I am still grateful for this positive attitude that triggered the onset of my career in scientific research. This challenging journey wouldn't have been possible without the strong support that I have received from friends, family and colleagues. First, I would like to thank my main supervisor Prof. Eivind Almaas. Your push for progress combined with excellent decisions, suggestions and feedback have ensured the completion of this PhD work. Furthermore, your encouragement to participation in international conferences and a research stay in Boston has been important: these exposures have been extremely motivating and instructive. Secondly, I would like to thank my co-supervisors at SINTEF, Håvard Sletta and Alexander Wentzel, for your consistent support and for sharing your broader knowledge on biotechnology.

In addition my three formal supervisors, I have also received invaluable supervision from Eduard Kerkhoven at Chalmers University and Prof. Daniel Segrè at Boston University. The six months long stay in Boston was one of the highlights of these four years, both because of Prof. Daniel Segrè's enthusiasm and motivating supervision, due to the many excellent group members. Of these, I would in particular thank David Bernstein and Shany Ofaim for excellent collaborations on the review of uncertainty in genome-scale models and for modelling of *Prochlorococcus*, respectively. I would also like to thank all group members of AlmaasLab at NTNU for good discussions and collaborations, and all colleagues at the Group of Biotechnology at SINTEF Industry for exposing me to all kinds of biotechnology applications and for your patience with my lack of lab experience. I would in particular like to mention Tjaša Kumelj whom I collaborated closely with during the two first years.

I would like to thank my closest family for paving the way towards academic research. Finally, I would like to thank my wife, Martine, not only for your patience when I'm too late home from work, but also for making sure that I have a life outside the office. I'm looking forward to new small and large adventures with you now that this doctoral thesis is completed.

Contents

Abstract	vi
Acknowledgement	vi
List of Papers	x
List of Abbreviations	xi
1 Introduction	1
2 Constraint-based flux analysis	5
2.1 Metabolic networks	5
2.2 Flux balance analysis	7
2.3 Unbiased analysis of the solution space	10
3 Reconstruction of genome-scale metabolic models	13
3.1 Genome annotation	13
3.2 Growth environment	14
3.3 Biomass	16
3.4 Gap-filling	16
3.5 Manual curation, evaluation and maintenance	17

4	GEM applications	19
4.1	Metabolic engineering	21
4.2	Omics integration	24
5	Dynamic flux balance analysis	31
6	Biosynthetic gene clusters	35
7	Summary of papers	39
8	Conclusion	43
9	Outlook	49
10	Bibliography	51
11	Paper 1	69
12	Paper 2	93
13	Paper 3	105
14	Paper 4	137
15	Paper 5	207

List of Papers

Included in the thesis

- 1. Addressing Uncertainty in Genome-Scale Metabolic Model Reconstruction and Analysis**
David B Bernstein*, Snorre Sulheim*, Eivind Almaas and Daniel Segrè.
Genome Biology 22, 64 (2021).
- 2. Predicting Strain Engineering Strategies Using iKS1317: A Genome-Scale Metabolic Model of *Streptomyces coelicolor***
Tjaša Kumelj*, Snorre Sulheim*, Alexander Wentzel and Eivind Almaas.
Biotechnology journal, 14(4), 1800180 (2018).
- 3. Automatic reconstruction of metabolic pathways from identified biosynthetic gene clusters**
Snorre Sulheim, Fredrik A Fossheim, Alexander Wentzel and Eivind Almaas.
BMC Bioinformatics 22, 81 (2021).
- 4. Enzyme-Constrained Models and Omics Analysis of *Streptomyces coelicolor* Reveal Metabolic Changes that Enhance Heterologous Production**
Snorre Sulheim, Tjaša Kumelj, Dino van Dissel, Ali Salehzadeh-Yazdi, Chao Du, Gilles P. van Wezel, Kay Nieselt, Alexander Wentzel, Eivind Almaas and Eduard Kerkhoven.
iScience, 23(9), 101525 (2020).

5. **Dynamic allocation of carbon storage and nutrient-dependent exudation in a revised genome-scale model of *Prochlorococcus***
Shany Ofaim*, Snorre Sulheim*, Eivind Almaas, Daniel Sher and Daniel Segrè.

Frontiers in Genetics, 12, 91 (2021).

* Equal contribution.

Not included in the thesis

1. **Computation Of Microbial Ecosystems in Time and Space (COMETS): An open source collaborative platform for modeling ecosystems metabolism**

Dukovski, Djordje Bajić, Jeremy M Chacón, Michael Quintin, Jean CC Vila, Snorre Sulheim, Alan R Pacheco, David B Bernstein, William J Rieh, Kirill S Korolev, Alvaro Sanchez, William R Harcombe and Daniel Segrè.

Under review at Nature Protocols. Preprint on arXiv.

List of Abbreviations

BGC	Biosynthetic gene cluster
BiGMeC	Biosynthetic Gene cluster Metabolic pathway Constructor
COBRA	Constraint-based reconstruction and analysis
dFBA	Dynamic flux balance analysis
FBA	Flux balance analysis
FVA	Flux variability analysis
GAM	Growth associated maintenance
GDLS	Genetic design through local search
GEM	Genome-scale metabolic model
MILP	Mixed integer linear programming
NGAM	Non-growth associated maintenance
NRP	Non-ribosomal peptide
NRPS	Non-ribosomal peptide synthetase
pFBA	Parsimonious flux balance analysis
PKS	Polyketide synthase
Sco-GEM	<i>Streptomyces coelicolor</i> GEM
TFA	Thermodynamics-based Flux Analysis

uFBA unsteady-FBA

Chapter 1

Introduction

Life is distinguished by the ability to convert simple molecules into complex structures ultimately leading to self-replication or reproduction. This feature is evident across all different forms of life, from small, and apparently simple bacteria, to complex multi-cellular eukaryotes such as plants or humans. Metabolism, the network of chemical reactions that occur in an organism, is one of the cornerstones that makes life alive and the focus of this doctoral thesis.

Cells and bacteria are to a large extent comprised of similar content. First we have the genome, that is the complete set of genetic material in an organism, made up of DNA. Most of the genome is found on one or several chromosomes tightly packed in the nucleus (in cells) or nucleoid (in bacteria), but DNA is also found on plasmids, in chloroplasts and in the mitochondria of cells. It is the genetic material of an organism that contains the specific information detailing growth, function and replication that is inherited by every daughter cell.

Genes are stretches of DNA in the genome that encode for the production a protein. The production of a protein from a gene is performed in several steps. First, the gene is transcribed to a single-stranded copy in the form of RNA by the enzyme RNA polymerase. If we now consider the somewhat simpler process in bacteria, this piece of RNA (called messenger RNA or mRNA) is translated according to the sequence of 3-letter codons into a sequence of amino acids provided by transport RNA and joined in the ribosome to make the final protein.

Proteins perform a range of different tasks including signalling, membrane transport and providing structure, but in the context of metabolism we are mostly interested in the proteins that act as enzymes. Enzymes catalyze chemical reactions and enable these reactions to occur at a rate that is sufficient to support cellular

maintenance, growth and replication. Furthermore, enzymes allow thermodynamically unfavourable reactions to occur through coupling with reactions that release energy (such as hydrolysis of ATP). Thus, it is the complete set of enzymes as dictated by the enzyme-encoding genes in the genome that determine the metabolic repertoire of a cell or bacterium, thereby defining the metabolic network of that species.

We have in the last four decades seen an unimaginable revolution in DNA sequencing technology (van Dijk et al. 2018) and cost¹ that allows us to sequence and assemble the complete genome of most organisms. For example, the Human Genome Project took 13 years from 1990 to complete one reference human genome (Collins et al. 2003), while the ongoing Earth BioGenome project aims to sequence all known eukaryotic species (1.5 millions) within 10 years (Lewin et al. 2018). With a complete genome at hand one can use bioinformatic tools (Aziz et al. 2008, Cantarel et al. 2008) to identify and functionally annotate genes in the genome (at least for bacteria, there is still a considerable error-rate for eukaryotes (Salzberg 2019)), and map out the organism's metabolism by connecting this information with reaction databases that detail the chemical reaction catalyzed by each enzyme. These reactions form the basis for genome-scale metabolic model (GEM) reconstruction, a process that is further discussed in Chapter 3.

A GEM is not only an organism-specific knowledgebase that can aid mapping or interpretation of big data, but when combined with linear algebra and optimization algorithms it functions as a mathematical framework that can predict metabolic phenotypes. This particular field of research, formally known as constraint-based reconstruction and analysis (COBRA), has seen a wide range of applications ranging from microbial communities to human medicine (Gu et al. 2019). The appreciation of this framework comes from its ability to encompass the complete, and extremely interconnected network of reactions that makes up metabolism, and by doing so enlighten non-intuitive connections that would otherwise be missed.

This thesis describes the reconstruction and use of GEMs, but also a review of the uncertainties that accompany this framework. We focus on two different species and applications, but the modelling framework provides a common thread throughout the text.

The first species, *Streptomyces coelicolor*, is a soil-dwelling bacterium armed with a huge set of genes responsible for the production of bioactive molecules (Bentley et al. 2002), and a model species for the phylum Actinobacteria that represents a major source of novel drugs (Berdy 2005). The work on *S. coelicolor* is con-

¹<https://www.genome.gov/sequencingcostsdata>

ducted within the INBioPharm² project where we aim to produce novel natural products through heterologous expression of biosynthetic gene clusters (BGCs) in *S. coelicolor*. Heterologous expression (i.e. transcription and translation of non-native genes) is one strategy for novel drug development that tries to overcome the issue that most BGCs are not expressed in their native host in standard lab cultivations (Rutledge and Challis 2015). Paper 2 and 4 contribute to the overall goal of this project by assessing potential strain design for enhanced performance as an expression host and elucidating metabolic characteristics of a strain previously engineered for that same purpose (Gomez-Escribano and Bibb 2011). Furthermore, both papers improve on existing genome-scale metabolic models of *S. coelicolor*, and as such provide a valuable tool both for the work presented here and for future research on this organism. We also predict strain-engineering strategies for *S. coelicolor* in Paper 3, but the main deliverable of this paper is a generic pipeline for the reconstruction of metabolic pathways encoded by BGCs.

The second species is a tiny, marine picocyanobacterium that performs photosynthesis, i.e. it uses sunlight, water and CO₂ to obtain its carbon and energy necessary for growth. Oxygen is a byproduct of this process. Another byproduct is organic material which contributes to the very bottom of the marine food web. Being the most abundant phototrophic marine bacterium, *Prochlorococcus* is responsible for about 3.9% of the global net primary production (Flombaum et al. 2013, Field et al. 1998), that is the amount of fixed CO₂ minus autotrophic respiration (Woodward 2007), and it has therefore a large impact on global climate. In Paper 5 we update and apply a genome-scale metabolic model of a specific ecotype, namely *Prochlorococcus marinus* MED4³, to understand how different environmental factors and the diel day-night cycle affects storage and release of fixed carbon.

The remaining parts of this doctoral thesis are outlined as follows: Chapter 2 covers basics of flux balance analysis (FBA) and associated methods used to analyse GEMs. Chapter 3 outlines the process of GEM reconstruction. These two topics are also discussed in depth in Paper 1, and these chapters therefore focus on basic knowledge that is omitted or only briefly discussed later. Chapter 4 describes GEM applications, however with a focus on metabolic engineering and omics integration which are of the most relevance to the content in Paper 2, 3 and 4. Chapter 5 describes an extension of FBA that is suited to model time-dependent dynamics, which is later applied in Paper 5. Chapter 6 presents a very brief primer on biosynthetic gene clusters, focused on the two biosynthetic classes addressed in Paper 3, namely polyketide synthases (PKSs) and non-ribosomal peptide syn-

²Integrated Novel Natural Product Discovery and Production Platform for Accelerated Biopharmaceutical Innovation from Microbial Biodiversity

³Now formally named *Prochlorococcus marinus* subsp. *pastoris* str. CMP1986.

theses (NRPSs). Collectively, Chapter 2-6 are intended to provide the background knowledge required to fully appreciate the detail later presented in the papers. However, within these chapters I also point to specific methods used in the papers where appropriate. Furthermore, in Chapter 4.2, covering omics integration, I briefly describe currently unpublished work on human alveolar macrophages. Chapter 7 summarizes each of the 5 papers, while a conclusion and a broader outlook are provided in Chapter 8 and 9, respectively. Finally, Paper 1-5, including any supplemental text and figures, are appended. For other supplemental material such as spreadsheets, BLAST-results etc., we refer the reader to the online material provided with the preprints / journal publications.

Chapter 2

Constraint-based flux analysis

Although mathematical models of metabolism already were put to use in biochemistry and biochemical engineering more than three decades ago (Fell and Small 1986, Tyson and Othmer 1978), it was the sequencing and assembly of whole microbial genomes (Blattner et al. 1997, Fleischmann et al. 1995) that enabled development of models on the scale of the genome (Edwards and Palsson 1999; 2000). During the last two decades the field has developed to cover a wide scope of applications and organisms, including bacteria, archaea, fungi, plants and mammals (Gu et al. 2019). Despite this variety, these models and their use share certain fundamental properties and assumptions that define the field of constraint-based reconstruction and analysis (COBRA). This basis is detailed below, starting with the representation of reactions and metabolites as a stoichiometric matrix and transitioning into flux balance analysis and other methods used to analyse these models. This provides the necessary background for Paper 1 which reviews sources of uncertainty, approaches for reducing uncertainty and frameworks that account for the uncertainty in model reconstruction and analysis.

2.1 Metabolic networks

A metabolic network is a formal representation of a set of biochemical transformations (in the remaining text referred to as reactions), where each node is a metabolite and each edge (link) represents a reaction. For example, the 3 reactions given in Equation (2.1) are illustrated as a metabolic network in Figure 2.1. Within a cell (or bacteria) most of the reactions are catalyzed by enzymes, but there are also spontaneous reactions. The rates at which these chemical transformations occur are in the following referred to as reaction fluxes, and within the COBRA field flux values are usually reported in mmol per gram cell dry weight per hour

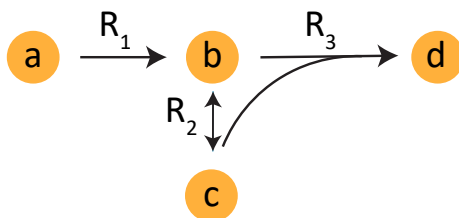


Figure 2.1: A minimal reaction network where the orange circles represent metabolites and the arrows represent reactions.

($\text{mmol gDW}^{-1} \text{ h}^{-1}$). Depending on the thermodynamic properties of each reaction and the concentration of each metabolite, reactions can either be reversible (like R_2) or irreversible, meaning that they can only proceed in one direction. While this network representation is sufficient in itself to understand the mass balance of each metabolite, or to grasp the effect of any network perturbation, a mathematical framework quickly becomes necessary with an increasing number of reactions. Genome-scale metabolic networks consist of up to several thousands of reactions and metabolites.



The stoichiometric matrix forms the basis for the mathematical framework used to analyse GEMs, and it is a matrix where each metabolite is represented by a row, and each reaction is represented by a column. Each matrix value reflects the stoichiometry of one metabolite in one reaction, with production and consumption represented as positive and negative values, respectively. For the example network in Figure 2.1 with 3 reactions and 4 metabolites, the corresponding stoichiometric matrix (\mathbf{S}), with 4 rows (a-d) and 3 columns (R_1 - R_3), is

$$\mathbf{S} = \begin{bmatrix} -1 & 0 & 0 \\ 1 & -1 & -1 \\ 0 & 2 & -2 \\ 0 & 0 & 1 \end{bmatrix}
 \tag{2.2}$$

However, in addition to intracellular reactions, GEMs also include transport reactions that facilitate transport across membranes, reactions that represent a particu-

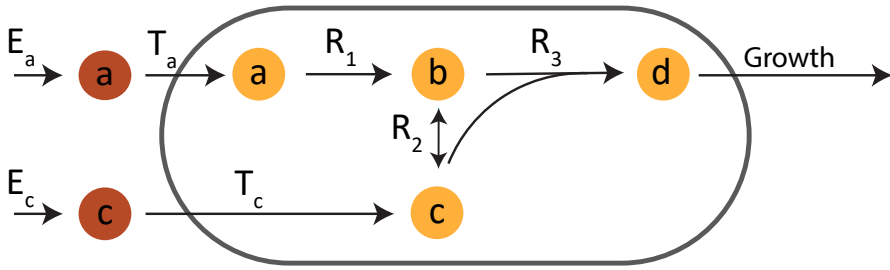


Figure 2.2: A minimal metabolic model where extracellular and intracellular metabolites are represented by red and yellow circles, respectively. The reactions are divided into exchange (E), transport (T), growth and normal reactions (R).

lar cellular process (such as growth) or boundary reactions that allow metabolites to enter or leave the system. By including these reactions, we get a minimal functional metabolic model (Figure 2.2).

2.2 Flux balance analysis

One important property of the stoichiometric matrix is that the mass balance of each metabolite (x_i) is the product of the matrix and the rate (flux) of each reaction (v_j), i.e.

$$\frac{dx_i}{dt} = \sum_j S_{ij}v_j, \quad (2.3)$$

or for all metabolites, written in matrix notation:

$$\mathbf{S}\mathbf{v} = \frac{d\mathbf{x}}{dt}. \quad (2.4)$$

This set of ordinary differential equations (ODEs) can be solved for short pathways where the number of reactions are limited. However, the rate laws that connect metabolite concentrations to reaction rates require numerical values for many kinetic parameters that are usually difficult to obtain. A further discussion on kinetic modelling is outside the scope of this work, but interested readers are pointed to the review by [Saa and Nielsen \(2017\)](#).

The ability to use this mathematical framework for metabolic models on the genome-scale comes from the assumption that the metabolism operates at different pseudo-steady states where metabolite levels are constant, i.e. no metabolite is accumulated or depleted. This assumption is reasoned based on the different time-scales of

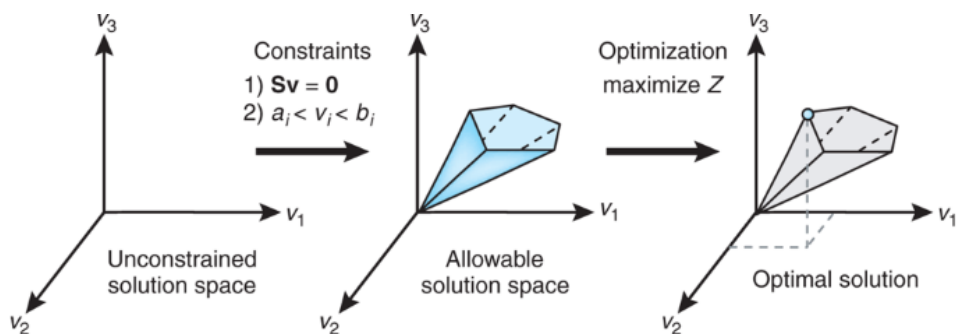


Figure 2.3: Illustrating how the steady-state assumption and variable constraints lead to a defined solution space that includes all allowable solutions. Flux balance analysis finds an optimal solution by maximizing or minimizing a chosen objective. Figure reused with permission from [Orth et al. 2010](#). Copyright 2010, Springer Nature.

cellular processes: The turnover of metabolites is on the order of seconds ([Buchholz et al. 2002](#)), much faster than the transcription, translation and degradation of proteins (minutes-hours) ([Maier et al. 2011](#), [Shamir et al. 2016](#)). The result of this assumption is that the right-hand side of Equation (2.4) is zero, reducing the set of ODEs to the algebraic Equation (2.5).

$$\mathbf{S}\mathbf{v} = \mathbf{0} \quad (2.5)$$

Solutions to this equation represent different combinations of reaction fluxes that all satisfy the mass balance constraint. Because the number of reactions usually is larger than the number of metabolites (or more precisely because the rank of \mathbf{S} is less than the number of unknown reaction fluxes) there is an infinite number of solutions to this equation. The n -dimensional volume spanned by these solutions is called the solution space (or the null space of \mathbf{S}). In addition to the steady-state constraint, solutions are further constrained by upper and lower bounds on the reaction fluxes. It is common practice to set the minimum/maximum rate of all reaction v_i to $-1000 \text{ mmol gDW}^{-1} \text{ h}^{-1} \leq v_i \leq 1000 \text{ mmol gDW}^{-1} \text{ h}^{-1}$. Although all reactions in principle are reversible, thermodynamic properties imply that certain reactions in practice only occur in one direction at cellular conditions. The direction of these reactions are then imposed as constraints by setting the upper or lower bound to zero. Furthermore, reaction bounds are used to constrain known (or estimated) intracellular fluxes (such as ATP maintenance) or define maximal uptake rates of metabolites present in the simulated nutrient environment. The applied reaction bounds transform the initially unconstrained solution space to a convex polytope that defines the allowable solution space (Figure 2.3).

Although the concept of a solution space might seem inconvenient at first, it is

actually a flexible framework that allows a range of analyses and data integration. First, one should appreciate that this solution space defines the range of possible metabolic phenotypes for a particular organism in a particular environment, and as such it defines to what extent the organism can act or adapt to reach a particular objective through cellular regulation or adaptive laboratory evolution (Ibarra et al. 2002). Flux Balance Analysis (FBA) is the most common method used in this context. FBA uses linear programming (also known as linear optimization) to find one solution within the allowable solution space that maximizes (or minimizes) a chosen objective (Orth et al. 2010, Varma and Palsson 1994a, Papoutsakis and Meyer 1985, Fell and Small 1986, Edwards et al. 2002). Therefore, any solution calculated by FBA is on the boundary of the solution space (Figure 2.3), either in a corner, on an edge or on an n-dimensional plane. If the direction of the objective function is perpendicular to any of the edges/planes that defines the solution space, the FBA solution is not unique because any point along this edge/plane has the same objective value. For GEMs, this is often the case. The FBA linear program is defined as:

$$\begin{aligned}
 & \text{maximize } Z = \mathbf{c}^T \mathbf{v} \\
 & \text{subject to :} \\
 & \quad \mathbf{S}\mathbf{v} = \mathbf{0}, \\
 & \quad a_i \leq v_i \leq b_i \quad \forall v_i \in \mathbf{v}.
 \end{aligned} \tag{2.6}$$

Here, we find a flux vector \mathbf{v} that maximizes the objective function \mathbf{c} . The lower and upper bounds for each reaction are given by a_i and b_i , respectively. Maximization of growth is the most common objective used in FBA (Gianchandani et al. 2008, Schuetz et al. 2007), based on the assumption that evolution has driven bacteria towards this maximum. Growth is represented in a GEM by a pseudo-reaction consuming all biomass components (Lachance et al. 2019). These components and their ratio can be experimentally determined (Beck et al. 2018). The applicability of growth maximization and other objectives is discussed in Paper 1. If we now continue to use the minimal cell example, we can use cobrapy (Ebrahim et al. 2013) to run FBA¹ and predict the maximal growth rate (the only biomass component is metabolite d) when both metabolite a and b are available in the growth medium with maximum uptake rates of 1 and 3 mmol gDW⁻¹ h⁻¹, respectively (Figure 2.4).

Many methods extend on FBA to capture more biologically relevant solutions, either with additional constraints or additional objectives (Lewis et al. 2012). Of

¹Get the notebook used to run this FBA calculation: <http://tiny.cc/minimal-cell-fba>

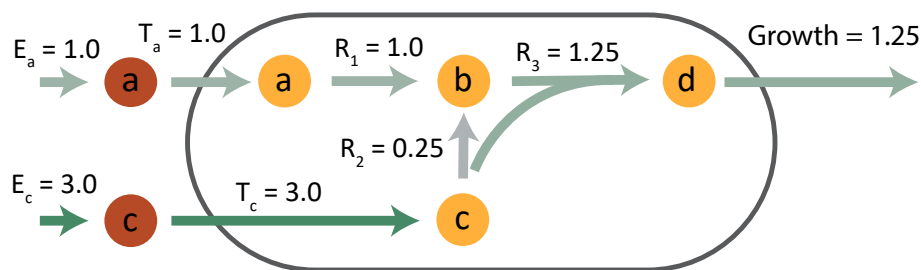


Figure 2.4: Predicted flux maximizing the growth (consumption of metabolite d) as predicted by FBA. The values are given in $\text{mmol gDW}^{-1} \text{h}^{-1}$ and the color of the reaction arrows corresponds to the flux rate.

these, parsimonious FBA (pFBA) is in particular worth mentioning because it provides a simple, but effective option that increases the accuracy of FBA by introducing a second optimization step minimizing the total sum of absolute reaction flux values (Lewis et al. 2010). The method is based on the idea that an organism strives to achieve its goal at a minimum enzymatic cost. Later work has provided some nuance to this picture by showing that microbes balance a trade-off between optimization towards one particular objective and the ability to quickly adapt to changing environments (Schuetz et al. 2012). Although pFBA doesn't account for differences in cellular cost of each enzyme nor the difference in enzymatic activity, the method has proven to be equally good or better than a range of more computationally demanding methods that leverage transcriptome data (Machado and Herrgård 2014). Methods that include the cost and efficiency of enzymes are discussed in Chapter 4.2, and in Paper 4 we use one of these methods to integrate time-series proteomics into a GEM to compare the engineered *S. coelicolor* strain M1152 with its wild-type ancestor.

2.3 Unbiased analysis of the solution space

Unbiased methods characterize the solution space without defining a cellular objective, and these methods are useful in situations where one cannot reasonably assume that the organism is striving towards a well-defined objective. Also, these methods can capture the variability of each reaction flux, thereby providing a much more comprehensive description of the metabolic phenotype compared to the single solution obtained with FBA or pFBA. Flux Variability Analysis (FVA) determines the possible flux range for each reaction (Mahadevan and Schilling 2003, Gudmundsson and Thiele 2010) in a given setting (e.g. in a given environment or at a determined growth rate), and is frequently used to analyze alternative optimal or suboptimal solutions (Reed and Palsson 2004), or to describe the

reduction in feasible phenotypes obtained from the introduction of additional constraints (Sánchez et al. 2017). Illustrations that contrast FVA with FBA and other approaches are provided in Figure 2.5.

There are more sophisticated analytical methods that can characterize or decompose the *optimal* solution space into biologically relevant elements (e.g. metabolic pathways) (Klamt et al. 2017, Schuster and Hilgetag 1994, Maarleveld et al. 2015). However, a characterization of the complete solution space for a normal sized GEM is for these methods still computationally intractable (Ullah et al. 2019). To this end, Monte-Carlo based sampling methods provide a feasible approach (Wiback et al. 2004). Most of the relevant methods rely on the hit-and-run approach (Almaas et al. 2004, Schellenberger and Palsson 2009) (Figure 2.5C), and recent implementations have drastically improved the sampling efficiency by artificial centering, scaling and rounding of the solution space, and removal of thermodynamically infeasible loops (Haraldsdóttir et al. 2017, De Martino et al. 2015, Megchelenbrink et al. 2014, Saa and Nielsen 2016). In Paper 4 we use a different sampling approach that samples the vertices of the solutions space (Figure 2.5D). This method does not suffer from auto-correlation between consecutive samples and is also supposed to provide a more realistic estimate of mean flux standard deviations (Bordel et al. 2010), but the method cannot provide the the same level of detail on flux distributions across the interior space as hit and run based methods. In Paper 1 we provide more detail on flux sampling approaches, and how they can be used to describe the uncertainty in model predictions.

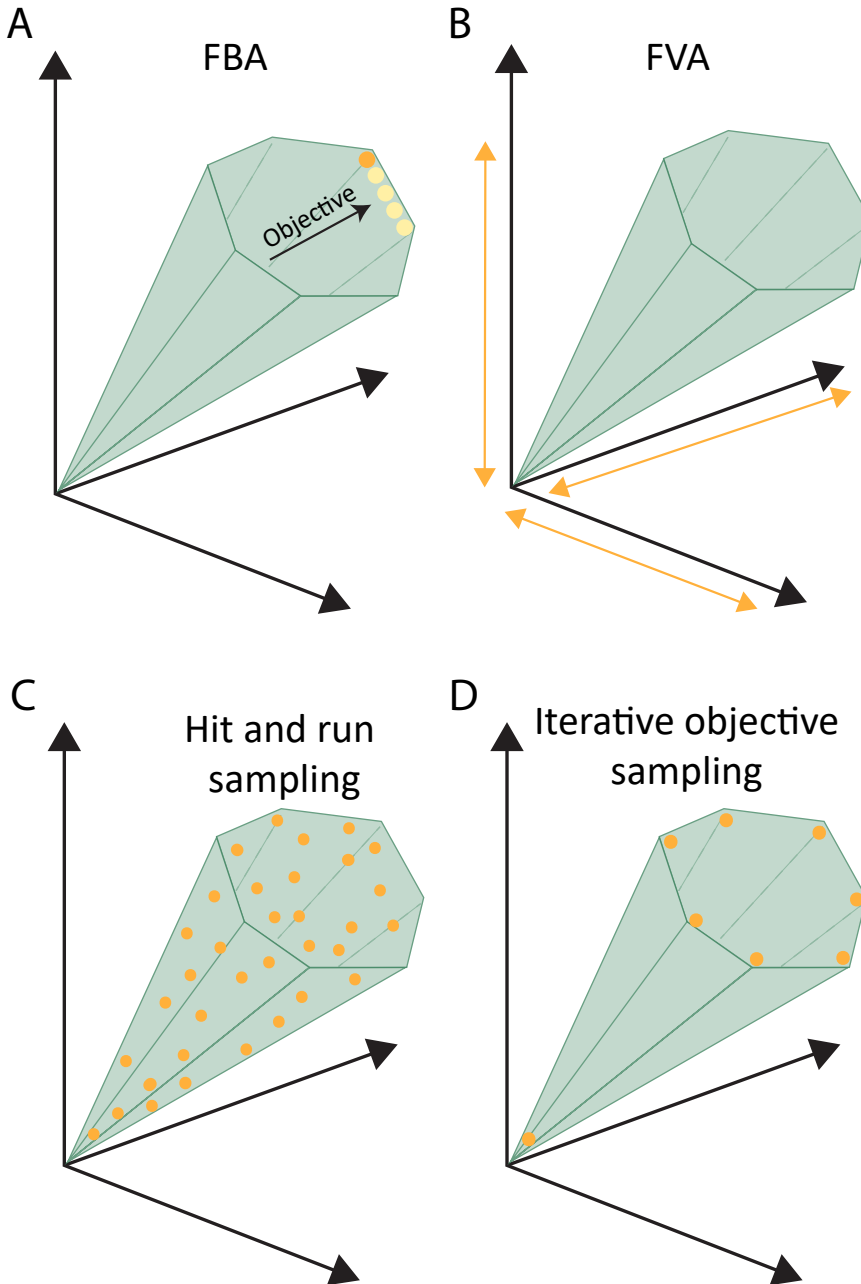


Figure 2.5: Comparison of FBA, FVA and flux sampling. A) FBA finds one objective solution (orange dot), however this solution is rarely unique (alternative solutions shown as light yellow dots). B) FVA identifies the maximum range for each flux variable. C) Hit and run approaches randomly sample flux distributions from the complete solutions space. D) The iterative objective sampling approach samples vertices of the solution space.

Chapter 3

Reconstruction of genome-scale metabolic models

Model reconstruction is the process of assembling organism specific information into a network representation that can be used to simulate metabolic phenotypes. All details are described in an available protocol ([Thiele and Palsson 2010](#)), but the process can in general be divided into 4 parts: 1) Genome annotation and assembly of a draft reaction network; 2) Specification of growth environment(s); 3) Description of the biomass composition; and 4) Gap-filling of draft network reconstruction. These four parts, and in particular the associated uncertainty, and approaches to mitigate, reduce or handle these uncertainties, are thoroughly discussed in Paper 1. Therefore, to avoid excessive repetition of content, this chapter provides a fairly minimal description of these four parts. The last part of this chapter covers manual curation and evaluation of model performance briefly. In combination with Paper 1, this content provides the necessary background for appreciating the model reconstruction efforts in Paper 2, 4 and 5.

3.1 Genome annotation

Annotation of a sequenced and assembled genome is the foundation for all model reconstruction efforts and comprises identification of all protein encoding genes in the genome and assignment of their function (Figure 3.1). For the purpose of drafting a metabolic network, we are primarily interested in enzyme-encoding genes and reactions catalyzed by the corresponding enzymes, in addition to membrane bound proteins that facilitate the transport of metabolites in and out of the cell or between different cellular compartments. The annotation of genes is in general based on sequence homology to genes that encode for an enzyme with known

function, and for prokaryotes usually performed with either RAST (Aziz et al. 2008) or PROKKA (Seemann 2014). Annotation of eukaryotic genomes are more complicated and apparently still challenging (Salzberg 2019). However, as we discuss in Paper 1, considerable improvement in genome annotation is also possible for prokaryotes by taking additional information into account.

Once the genes are annotated to specific enzymes one can map these genes to reaction databases such as KEGG (Kanehisa and Goto 2000), MetaCyc (Karp et al. 2002) or MetaNetX (Moretti et al. 2016). This mapping does not only link genes to reactions (via enzymes or transport proteins), but based on information in these databases one obtains details about each reaction (full name, database ID and related metabolic pathways) and all associated metabolites (full name, chemical formula, molecular weight and ionic charge). The gene-protein-reaction association can be complex: several genes can encode the same enzyme (isogenes), enzymes can be promiscuous and accept a variety of substrates, and thus facilitate several different reactions in the metabolic network, and some reactions are catalyzed by enzymatic complexes where all subunits are encoded by different genes. To cover all these options the gene-protein-reaction mapping follows a boolean logic based on *and* and *or* associations ensuring that *in silico* genetic modifications are correctly propagated to the metabolic network. The endpoint for the genome annotation and reaction mapping processes is a draft metabolic network that can be further curated manually or by gap-filling.

There are several pipelines that automate the generation of a draft metabolic network reconstruction, and these have been reviewed recently (Mendoza et al. 2019). These pipelines drastically speed up the reconstruction process, however additional manual curation and quality control are still necessary to obtain a high-quality model.

3.2 Growth environment

Both for subsequent gap-filling, and to simulate metabolic phenotypes in general, it is necessary to determine at least one nutritional environment that supports growth of the species subject to model reconstruction. As it is customary to assume that all metal ions are present, this usually boils down to determine anaerobic or aerobic conditions, available carbon, nitrogen, phosphate and sulphur containing compounds, and at what rates these compounds can be consumed by the organism. However, certain species require additional environmental stimuli to grow, e.g. light for the phototrophic *Prochlorococcus* which is implemented and discussed in Paper 5.

It is relatively easy to define growth-supporting environments for well-described

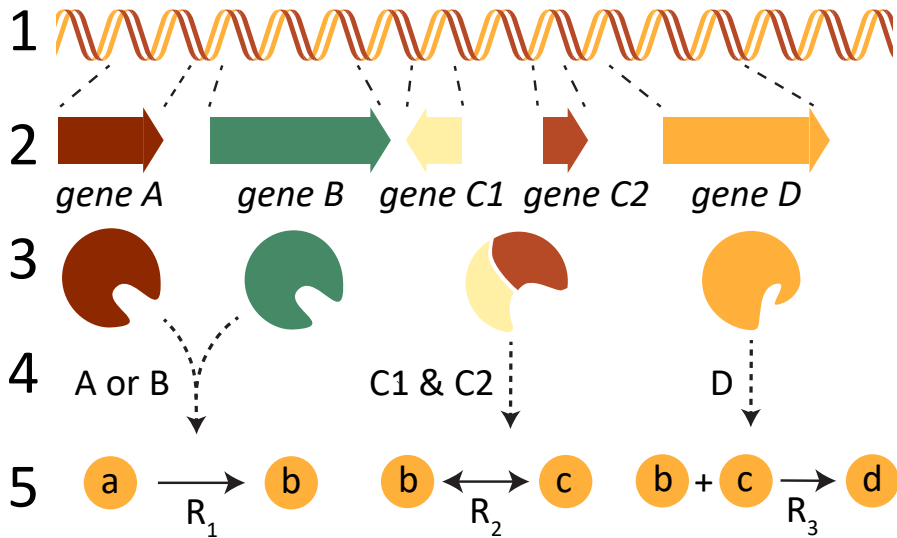


Figure 3.1: The reactions (5) used to create a draft network are obtained from the genome sequence (1) in several steps. First, annotation of the genome provides (2) an overview of genes and (3) enzymes that the genes encode. Then, based on the function of each enzyme one generates (4) gene-protein-reaction mappings and (5) a first draft of the possible metabolic reactions.

organisms, either by searching the literature or by using public databases such as MediaDB (Richards et al. 2014) or KOMODO (Oberhardt et al. 2015). It is more challenging to accurately determine specific uptake rates, but these can be calculated from measured metabolite depletion of the respective medium components. A particular challenge in this context is that uptake rates can vary between different growth stages (as seen in Paper 4 for batch fermentation of *S. coelicolor*). Furthermore, microbes can both co-utilize nutrients or consume nutrients sequentially (known as diauxic growth). One can address these challenges by using different uptake rates for different growth stages, or by incorporating additional constraints on resource allocation (Salvy and Hatzimanikatis 2020). Nevertheless, for subsequent quality control of predicted growth rate, one should also measure the growth, either directly by measuring cell dry weight, by optical density, or by viable cell counts. Furthermore, measurements of secreted byproducts can provide valuable constraints that improve model predictions. When the required experimental data is in place, the calculations of growth, uptake and secretion rates are in principle straightforward. However, when studying the change in metabolism through different growth stages and medium depletion (such as for *S. coelicolor* in Paper

4), accurately estimating these rates can be challenging, both because of insufficient temporal resolution and unclear growth phase transitions. Dynamic FBA, where one simulates the time course of microbial growth, requires additional kinetic parameters that further complicates the determination of parameters associated with the growth environment.

3.3 Biomass

Within the COBRA framework, growth is represented by a model (biomass) reaction that consumes all compounds used to build the organisms' macromolecular components required for growth, weighted by their respective contribution. The major components are proteins, RNA, DNA, lipids and carbohydrates, but there is also a need for a range of vitamins and cofactors. The stoichiometric coefficients in this biomass reaction are scaled to represent the amount needed to make 1 gram of dry cell weight of the specific organism. Several recent laboratory protocols describe how these biomass components can be measured ([Beck et al. 2018](#), [Széliová et al. 2020](#)). Downstream calculations are now facilitated by a recent protocol and software ([Lachance et al. 2019](#)).

In addition to the biomass components, one must also estimate the basis energy required for cellular maintenance and the energy required for growth. The growth associated maintenance (GAM) energy is included in the biomass reaction, while the non-growth associated maintenance (NGAM) requirement is represented by an ATP hydrolysis pseudoreaction that converts ATP to ADP and phosphate. The GAM and NGAM requirements are estimated from the slope and intercept of the linear trendline fitted to substrate consumption data or ATP demand estimates at different growth rates ([Thiele and Palsson 2010](#), [Lachance et al. 2019](#)).

3.4 Gap-filling

The draft network reconstruction obtained from mapping the annotated genome to reaction databases has usually missing links that render the GEM infeasible (can't produce all biomass components) in one or several defined growth environments. In general, gap-filling algorithms use a reaction database to include additional reactions (without genomic evidence) to fulfill known growth phenotypes. The new reactions are usually selected with a MILP-approach that finds the minimal set of reactions required to fulfill one growth phenotype. The complete gap-filling procedure iteratively uses this MILP-approach for a number of different known growth phenotypes. However, the final reaction network depends on the order that the growth discrepancies are resolved ([Biggs and Papin 2017](#)), highlighting the need for a probabilistic or ensemble-based approach, as recently demonstrated by [Medlock and Papin \(2020\)](#) and further discussed in Paper 1.

3.5 Manual curation, evaluation and maintenance

Manual curation of the metabolic network represents a significant amount of the work required to make a high-quality GEM. In basic terms, this involves correcting errors introduced in the automatic draft reconstruction process, curate and include additional annotations to ensure interoperability, and curate the model network to improve the accuracy in predicting known phenotypes. This process is iterative, and often requires repeated cycles that introduce and evaluate the effect of new changes.

A large library of growth and gene knockout phenotypes is very valuable in the evaluation of model performance, and in particular large-scale transposon mutagenesis or CRISPR knock-out studies that identify essential and non-essential genes. Note that this type of knockout screens may not differentiate between lethal gene knockouts and very slow growing mutants, and in Paper 2 we use a 50% reduction in growth rate as a threshold when comparing *in silico* knockout phenotypes with the transposon mutagenesis data.

One issue with automatic model reconstruction pipelines, or similar semi-automatic approaches used to search, map and obtain additional reactions databases, that was encountered during our development of the *S. coelicolor* GEM is the presence of redundant or very similar reactions and metabolites. E.g. KEGG contains for certain compounds both a generic form and stereoisomers (e.g. D-glucose¹). Enzymes might be selective for one specific stereoisomer, and in the biosynthesis of certain compounds different stereoisomers may lead to different products. For example, polymerization of α D-glucose yields starch while polymerization of β D-glucose yields cellulose. In terms of model reconstruction, ambiguity in reaction specificity and presence of multiple variants of the same compound can lead to redundant reactions or possibly disconnected pathways. Continuing with D-glucose as the example, if one naively includes all reactions in KEGG that map to the enzyme hexokinase (EC: 2.7.1.1) one obtains at least three reactions that converts D-glucose to glucose 6-phosphate, one for each of the three aforementioned variants.

It is important to ensure that all reactions in a metabolic model are mass and charge balanced, to avoid non-biological solutions where mass or charge appear from nothing. Because different reaction databases differ in how they report the charge and chemical formula of metabolites, this may provide a significant amount of work if one combine results from different databases. E.g KEGG reports the chemical formula of the uncharged metabolite, while BioCyc reports the chemical

¹D-glucose (C00031), α D-glucose (C00267) and β D-glucose (C00221).

formula of the charged metabolite ion. Metabolites often appear as ions intracellularly, and this is the preferred representation in GEMs. Despite the known importance of mass and charge balancing, this was not achievable within our development of the *S. coelicolor* GEM because of the extremely rich secondary metabolism, where many of the intermediate compounds are not sufficiently characterized.

Substantial annotation to different namespaces increases the value of the GEM, because it eases integration of different omics data or *in silico* co-cultivation of microbial species which require a common namespace. As a general recommendation, the MetaNetX cross reference seems to be the best link between different reaction databases. However, for naming of model entities the BiGG nomenclature (King et al. 2016) seems preferable because of human interpretable abbreviations.

The COBRA community would in general benefit from more clear community standards (Carey et al. 2020). An important contribution in this context is the MEMOTE software for assessing the quality and coverage of a GEM (Lieven et al. 2020). Furthermore, the MEMOTE test suite can easily be expanded to cover custom tests, as done in Sco-GEM for knockout and growth phenotypes. As such, the MEMOTE framework and report can be leveraged to track model development and do automatic unit testing in line the common practice in software development.

Chapter 4

GEM applications

The scope of genome-scale metabolic models has expanded drastically during the last decades, both because of increased data availability and increased diversity of model reconstructions that covers the complete range from relatively simple bacteria to male and female whole-body human models (Thiele et al. 2020). In combination with a massive development of methods and algorithm (Figure 4.1), this has allowed the application of GEMs to a wide range of industrial and scientific topics, from human medicine to microbial metabolism and ecology. A comprehensive description of all methods and applications is beyond the scope of this thesis, and we therefore point the reader to existing reviews (Gu et al. 2019, Lewis et al. 2012). Rather, we focus on metabolic engineering and integration of omics data, the two most relevant subtopics.

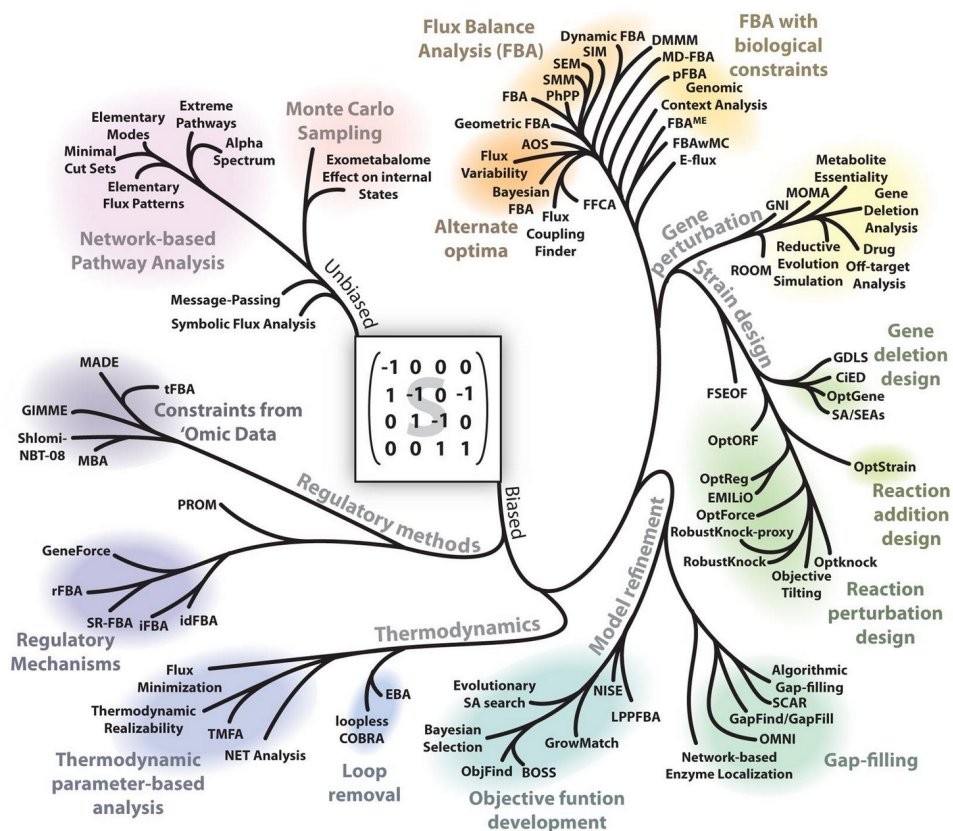


Figure 4.1: There is now a complete phylogeny of methods for the analysis of GEMs. These methods can be categorized as biased approaches (assuming that the organism has a specific cellular objective) or unbiased approaches (such as random sampling). Figure reused from [Lewis et al. 2012](#). Copyright 2010, Springer Nature.

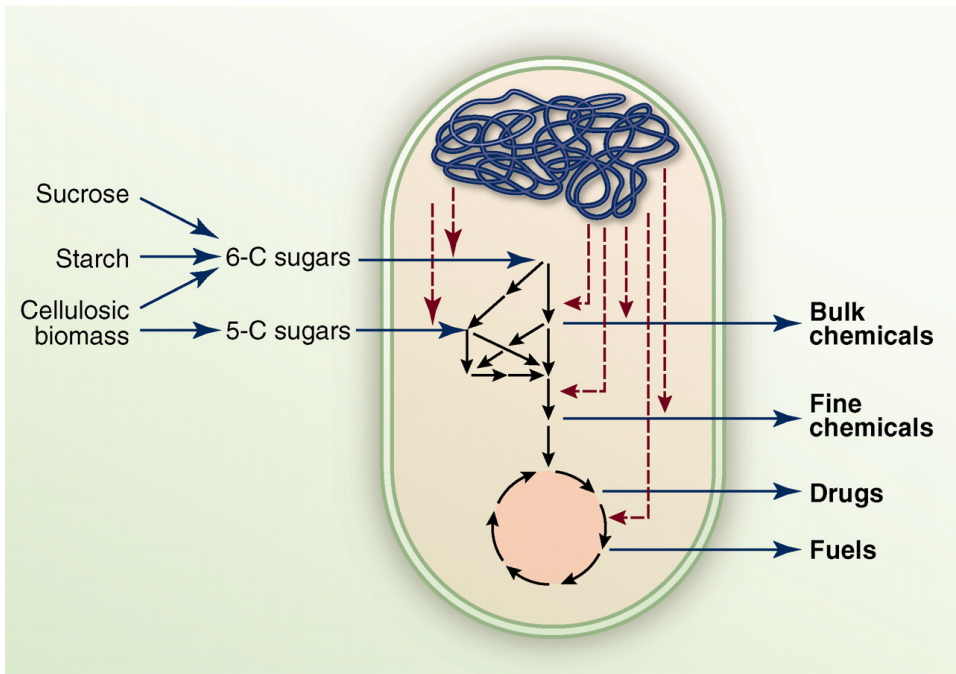


Figure 4.2: Illustrating the concept of cell factories. Figure reused from [Keasling 2010](#) with permission from AAAS.

4.1 Metabolic engineering

Metabolic engineering is the design of cells or microbes to improve their function as cell factories, producing valuable compounds such as drugs, biofuel and chemicals from renewable feedstocks (Figure 4.2, [Keasling \(2010\)](#), [Nielsen \(2001\)](#)).

Once the ability of GEMs to predict the effect of gene deletions was demonstrated ([Förster et al. 2003](#), [Edwards and Palsson 2000](#)), this quickly emerged as a potential application of genome-scale models [Patil et al. \(2004\)](#). Further work in this context showed that the growth phenotype of gene knockout mutants could be more accurately predicted by identifying the solution closest to the former wild-type solution while satisfying the additional constraint placed on the network by gene deletion ([Segre et al. 2002](#)).

Engineering of metabolic strains for a particular purpose was traditionally achieved by inducing random mutations and selecting optimal mutants by screening or adaptive laboratory evolution. However, these strategies give little, or no, knowledge about the mechanisms underlying the improved characteristics. With the

availability of genome-scale reconstruction of metabolic networks one could now suggest rational strain-engineering strategies based on mechanistic genotype-phenotype relations. To this end, Burgard and coworkers developed the first algorithms for predicting optimal gene additions (Burgard and Maranas 2001) and knockouts (Burgard et al. 2003). An important concept here is growth coupling. Here, one aims to make the production of the target compound an obligatory by-product of growth, so that the organism must produce this compound to achieve maximal growth rate (Figure 4.3 A). In the wake of these contributions, several different flavours addressing the same problem occurred (Figure 4.3B and C). RobustKnock (Tepper and Shlomi 2010) and objective function tilting (Feist et al. 2010) also identify optimal reaction knockout strategies, while OptGene (Patil et al. 2005) and Genetic Design through Local Search (Lun et al. 2009) find optimal gene knockout strategies, and are therefore more biologically relevant with respect to the effect of corresponding genetic modifications. The latter method, Genetic Design through Local Search, differs from the other methods by using heuristics combined with optimization to identify possible solutions. This, and other similar approaches, does not guarantee an optimal solution, but the computational cost scales better with the number of knockouts (Patil et al. 2004), and is therefore in practice required for more than three knockouts. Several of these methods have been experimentally validated, and later work have expanded the scope of predictions to up and down regulation of specific genes, multiple objectives and more (Maia et al. 2016).

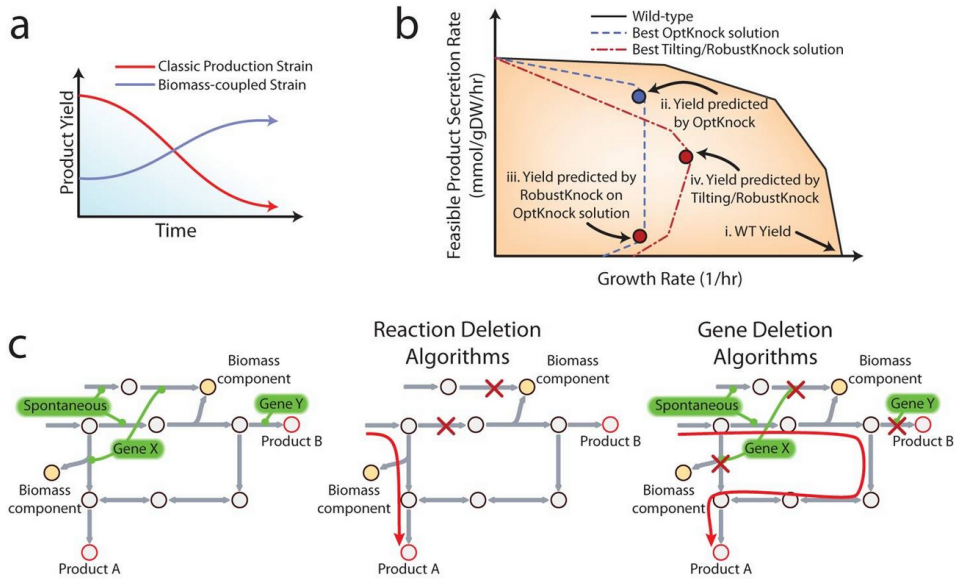


Figure 4.3: Different aspects of rational strain design: A) growth coupling; B) the production envelope and different solutions identified with different methods; C) Reaction vs gene deletion algorithms. Figure reused from [Lewis et al. 2012](#). Copyright 2010, Springer Nature.

4.2 Omics integration

Omics is a generic term that covers all kinds of large scale experimental measurements or data sets that describes one specific layer of cellular or microbial life. The following is not an exhaustive list of omic types, but it is sufficient in this context: Genomic tools provide sequence data that describe genes and genomes, transcriptomics measure RNA levels to detail the parts of the genome that are transcribed under a specific condition, and metabolomics describe the abundance of all intracellular metabolites. Technological developments have drastically increased the access to different omics data, hence the increased use of GEMs in this context. Genomics forms the basis for GEM reconstruction, and can furthermore be used in multistrain analysis (Fang et al. 2020) or to interpret the effect of genetic differences (Øyås et al. 2020). Below we focus on the other omic flavours that are more commonly integrated with GEMs to improve model predictions.

Transcriptomics

Transcriptomics can be obtained either from microarrays with gene and organisms specific probes, or more commonly nowadays, from RNA sequencing that determines the nucleotide sequence of all RNA in the sample (Stark et al. 2019). In contrast to other bioinformatic analyses, such as gene set enrichment analysis (Subramanian et al. 2005) and gene-co expression networks (Voigt et al. 2017), or data visualization methods such as clustering, dimension reduction techniques and data mapping (Gehlenborg et al. 2010), GEMs provide a mechanistic description of biological systems that can reveal causal relations. For example, several methods employ transcriptomics to constrain the solution space according to the transcript levels of the enzyme-encoding genes by modulating the allowed flux through the corresponding metabolic reactions as defined by the gene-protein-reaction associations. The fluxes can be constrained either in a continuous fashion (Colijn et al. 2009, Collins et al. 2012) or by turning metabolic reactions on/off according to a determined expression level threshold (Åkesson et al. 2004, Jensen and Papin 2011). Other methods omit the need of a defined cellular objective by rather minimizing the difference between the flux distribution and gene expression data (Zur et al. 2010, Lee et al. 2012, Shlomi et al. 2008). This concept is in particular useful in multicellular organisms or other situations where a clear cellular objective is not easily defined. However, when benchmarking these methods, none consistently outperforms the other methods nor parsimonious FBA which doesn't use gene expression data at all (Machado and Herrgård 2014). One may hypothesize that this lack of gain is due to immature or poorly developed methods, however it seems more likely that this lack of success derives from the fundamental limitations of this concept, where gene expression levels are used as a proxy for enzyme

abundances. Despite the clear mechanistic relation, several studies have shown varying correlations between these two biological entities (Waldbauer et al. 2012, Olivares-Hernández et al. 2011, Jayapal et al. 2008).

Other methods integrate gene expression data indirectly through transcriptional regulatory networks (TRNs) (Cruz et al. 2020, Faria et al. 2014). TRNs are networks where the nodes are regulators and target genes, and the link describes the regulatory association between these two elements. By integrating these networks one can account for regulatory mechanisms that would otherwise not be considered in FBA, and thereby improve model predictions (Cruz et al. 2020, Faria et al. 2014). However, because standardized frameworks for TRN reconstruction are lacking and the data required for their reconstruction are not readily available, complete TRNs, and therefore the use of these algorithms, are limited to the few, most well-described organisms (Cruz et al. 2020). Transcriptomics can also be used to generate context-specific models, representing e.g. a specific cell type in multicellular organisms (Agren et al. 2012) or different individuals (Agren et al. 2014). Their reconstruction is based on tailoring a generic GEM according to the genes that are expressed in each context. For a more detailed description of this methodology we refer the reader to a recent review (Cho et al. 2019).

Proteomics

One obvious limitation of FBA is the lack of constraints that ensure predicted fluxes to be within a range that is feasible given the amount and efficiency of the present enzymes. Although the reaction flux (v) of an enzymatic reaction, as described by the Michaelis-Menten equation (Equation 4.1, see Schnell (2014) for a discussion on assumptions and limitations), depends on the substrate concentration $[S]$, one can define a maximum possible rate as the product of the turnover rate and concentration of each enzyme, i.e. $V_{max} = [E] \cdot k_{cat}$. The turnover rates (k_{cat} values) can be obtained from enzyme databases such as BRENDA (Schomburg et al. 2002), however one should be aware that these values come with a considerable uncertainty, both because of different experimental conditions (Bar-Even et al. 2011) and a difference between *in vivo* and *in vitro* values (Davidi et al. 2016). Furthermore, the coefficients are often not available for all enzyme-substrate combinations or for the particular organism of interest. E.g. in our development of the enzyme-constrained *S. coelicolor* GEM in Paper 4, 88% of the obtained turnover rates matched the exact EC number, 32% matched the correct substrate, and only 5% of the values were measured in *S. coelicolor*.

$$v = V_{max} \frac{[S]}{K_M + [S]} \quad (4.1)$$

The first COBRA methods to incorporate enzyme coefficients used a global constraint to represent the limited cellular volume (Beg et al. 2007), limited membrane space (Zhuang et al. 2011b), macromolecular resource allocation (Goelzer et al. 2011), or maximum protein mass, either in total (Adadi et al. 2012) or divided by different cellular functions (Mori et al. 2016). These methods have indeed improved GEM predictions, e.g. by predicting inefficient metabolism in cancer cells (Shlomi et al. 2011) and overflow metabolism in *E. coli* (Basan et al. 2015). ME-models also include macromolecular allocation constraints, in addition to a more detailed description of the cellular machinery required to produce these macromolecules (Thiele et al. 2012, O'Brien et al. 2013). For an in depth review of resource allocation in GEMs we refer the reader to a recent review (Yang et al. 2018).

GECKO is a recently developed extension of the global protein constraint methods, formulated and designed to constrain the flux through each reaction according to measured protein abundances (Figure 4.4). In this framework, each reaction catalyzed by an enzyme is modified to include this enzyme as a pseudo-substrate. These pseudo substrates, that represent allocation of each enzyme, are drawn either from an enzyme pool (if the abundance of this enzyme is not measured) or from an exchange reaction with its flux constrained according to the measured abundance. The reconstruction of GECKO-formulated GEMs is carried out in a semi-automatic fashion: initial enzyme coefficients are automatically obtained from BRENDA (Schomburg et al. 2002), but these values must be curated to avoid over-constraining the model. This curation is performed by iteratively identifying and modifying growth limiting turnover rates until the model can sustain a growth rate as expected from experimental data. In Paper 4 we use this framework to generate in total 17 different GEMs according to strain and time point specific proteome data.

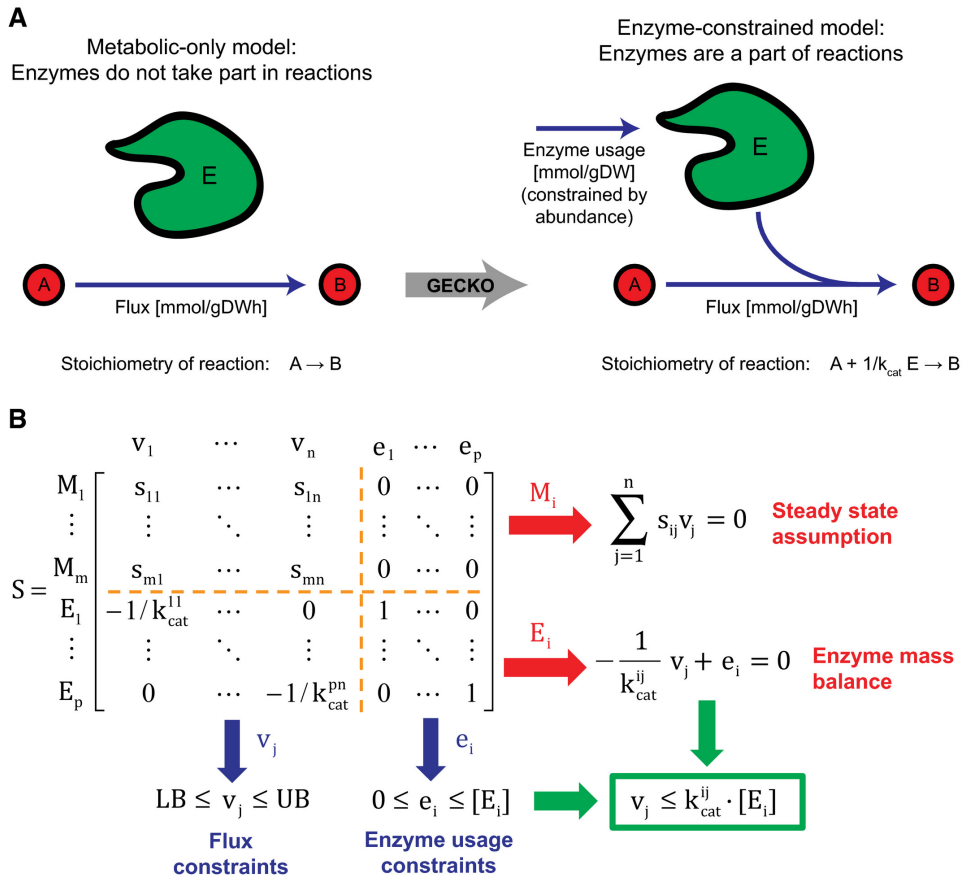


Figure 4.4: Illustration of the GECKO method for incorporation of enzyme coefficients and proteomics. **A)** The difference between a classical GEM and a GECKO-formulated extension that takes into account the allocation of enzymes required to maintain the predicted flux through each reaction. **B)** The incorporation of enzyme coefficients and proteomics is mathematically achieved by extending the stoichiometric by extra rows that describes the enzyme abundance constraints, limiting the allocation of each enzyme to the measured abundance. Figure reused from [Sánchez et al. \(2017\)](#), under the license [CC-BY-4.0](#)

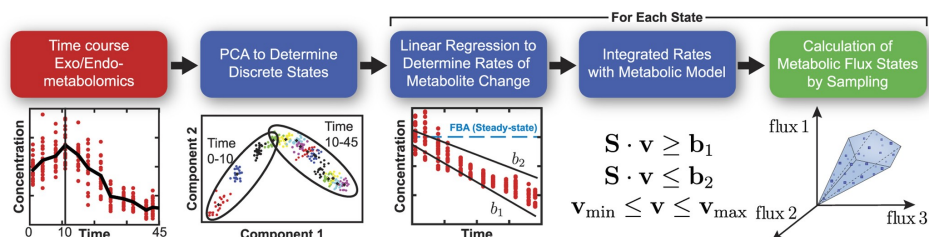


Figure 4.5: Illustration of the unsteady-FBA approach. Figure reused from [Bordbar et al. 2017](#), under the license [CC-BY-4.0](#).

Metabolomics

As metabolite concentrations are not represented in FBA, the use of (intracellular) metabolomics data requires additional computation for its integration. Note that exometabolomics, i.e. measurements of the concentration of growth medium components, are frequently used to estimate GEM uptake and secretion rates ([Aurich et al. 2016](#)), and therefore differ from intracellular metabolite measurements in its applicability. The unsteady-FBA (uFBA) approach represents one successful integration of metabolomics data (Figure 4.5; [Bordbar et al. 2017](#)). Here, the authors use time-course metabolite measurements to estimate the rate of change of both medium components and intracellular metabolites, and subsequently integrate metabolite accumulation or depletion into the right hand side of the mass balance equation. This gives an increased accuracy in the prediction of dynamic flux states on red blood cells, platelets and in *S. cerevisiae* (baker’s yeast).

Integration of metabolomics data in a human alveolar macrophage GEM

In currently unpublished work, we have tried to use uFBA to reveal metabolic differences between human alveolar macrophages stimulated by different ligands. A ligand is a molecule that binds to a certain receptor to trigger a cellular response, and in this work we used 7 different ligands that bind to Toll-like receptors to trigger an immune response similar to the one that occurs upon detection of a pathogenic microbe. The metabolomics data cover describe the concentration of 60 different intracellular metabolites measured at five time points, i.e. at 2, 4, 6, 8 and 24 hours. Prior to uFBA calculations, we used statistical analyses to identify metabolite concentrations that differed between the ligand-stimulated macrophages and the control. We found a few important metabolites, and in particular a range of nucleotides, that showed a clear difference across several ligands. However, for the uFBA integration, we first realized that very few metabolite accumulation or depletion rates were significantly different from zero. We interpreted this as a low signal-to-noise ratio associated with this type of measurements, rather than actu-

ally a lack of signal. We continued with the calculated trends and used uFBA to predict the metabolic state for each time interval for each of the 8 differently stimulated macrophage samples. We also used extracellular measurements of the glucose and glutamine consumption in addition to lactate production to constrain the models. However, these exometabolome measurements were only conducted at the end of the experiment, and thus, the uptake and secretion rates were constrained to the same values for all time intervals within each macrophage sample. We used a parsimonious variant of uFBA, both with nitric oxide production and ATP production as objectives to predict metabolic phenotypes, and found that the signal from the integrated intracellular metabolite accumulation or depletion rates was very small compared to the effect of different uptake and secretion rates. It was therefore difficult to obtain any time point specific details from these analyses, and we concluded that a shorter time scale (so that the accumulation and depletion rates are larger) are necessary for a successful use of uFBA.

Instead, we implemented another method for metabolomics data integration. Thermodynamic-based Flux Analysis (TFA) uses calculations of the Gibbs Free energy to ensure that all reactions are thermodynamically feasible according to the second law of thermodynamics ([Henry et al. 2007](#)), i.e. a reaction flux can only be positive if the change in Gibbs Free energy of that reaction is negative. The TFA method has been used to characterize different biological systems ([Stanway et al. 2019](#), [Hadadi et al. 2020](#)), but to our knowledge, not in such a large scale comparison of strains and time points. The change in Gibbs Free energy can be calculated with the group contribution method ([Jankowski et al. 2008](#)), and depends on the chemical groups and concentration of the reaction substrates and products. Therefore, our hypothesis is that the different metabolite concentrations affect the macrophage metabolism by changing the direction (or reversibility) of certain reactions. We have created TFA models for all time points and macrophage samples using pyTFA ([Salvy et al. 2019](#)), and are now analysing this data. This analysis has been partly hampered by lack of robust and reproducible results that may derive from the python implementation of TFA, but preliminary results indicate that some of the reactions in the nucleotide metabolism are indeed affected by the different metabolite levels.

Chapter 5

Dynamic flux balance analysis

We have up to this point not considered dynamic systems that require a temporal dimension in FBA. The study of *Prochlorococcus* is such a system, where the metabolism has a diel periodicity that is dictated by light-driven photosynthesis (Zinser et al. 2009). This leads to a dynamic accumulation and depletion of carbon storage in the form of glycogen (Szul et al. 2019).

Dynamic FBA (dFBA) is an extension of FBA that replicates the dynamics of growth by dividing the total time span into discrete time intervals (Varma and Pals-son 1994b, Mahadevan et al. 2002). FBA or pFBA can then be used to sequentially simulate growth in each time step (Figure 5.1). Between each time step, the FBA result is used to update the medium composition according to the metabolites that have been consumed or produced by the organism, and the total biomass according to the predicted growth rate. Thus, in contrast to regular FBA, where maximal uptake rates are explicitly defined as bounds on exchange reactions, the uptake rates in dFBA are predicted from the metabolite concentrations in the medium, usually using the Michaelis-Menten equation (Equation 4.1). Therefore, dFBA requires the additional parameters V_{max} and K_M for each nutrient to enable accurate predictions.

Several different flavours of dFBA have been used to study the diel cycle in cyanobacteria (Baroukh et al. 2014; 2015, Sarkar et al. 2019, Rügen et al. 2015, Reimers et al. 2017, Knoop et al. 2013), with different methods used to account for the varying light availability and hence ability to produce biomass components. For example, in the CycleSyn method each day is divided into 2-hour periods, each time period with different light availability (Sarkar et al. 2019). Furthermore, reaction fluxes are constrained by temporal transcriptomics data and intracellular

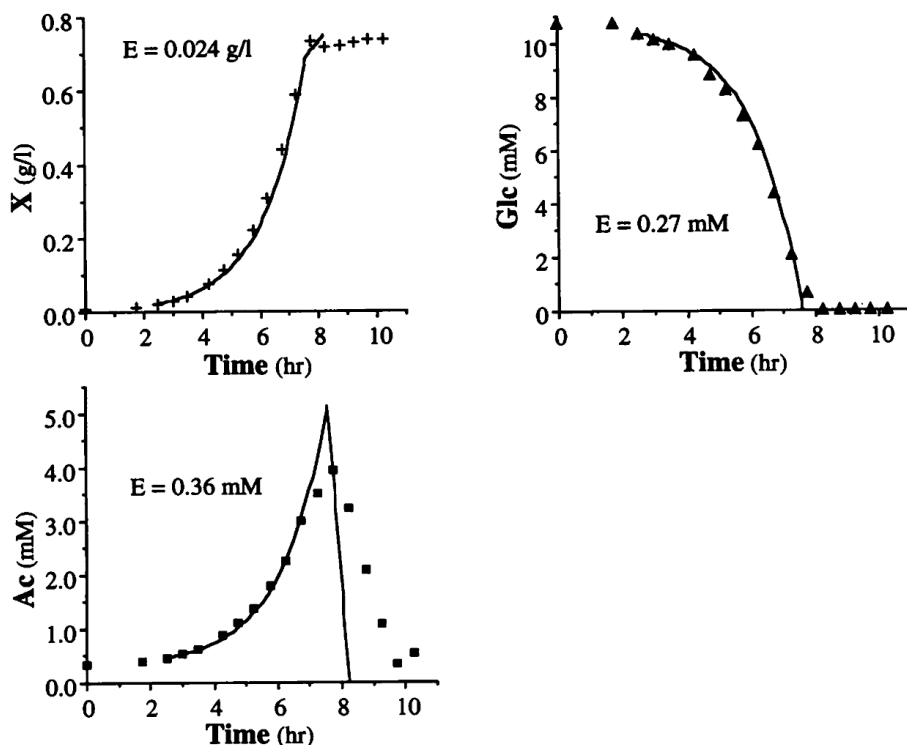


Figure 5.1: Illustration of Dynamic FBA, here used to simulate the batch fermentation of *E. coli*. The top left panel shows the growth, the top right panel shows glucose consumption, and the bottom panel shows first acetate production, followed by acetate consumption once all the glucose is consumed. Figure reused from [Varma and Palsson 1994b](#) with permission. Copyright 1994, American Society for Microbiology.

metabolites can be carried over to later time periods to allow accumulation of important nutrients during the day. Then, the reaction fluxes as well as transfer fluxes (representing storage) in all time periods are predicted in a single optimization step that optimizes the drain of biomass in the last time period. This allows different biomass components to be produced at different time points throughout the day, and was shown to simulate measured glycogen storage in *Synechocystis*. A second approach uses a standard dFBA scheme but imposes a time-dependent biomass equation ([Knoop et al. 2013](#)). Finally, a third, and very interesting approach that is similar to Resource Balance Analysis ([Goelzer et al. 2011](#)), takes into account the allocation of all macromolecules required to sustain light absorption, metabolism and growth throughout the diel cycle ([Rügen et al. 2015](#), [Reimers et al. 2017](#)). The approach uses a non-linear optimization to maximize the the total growth over 24

hours, and this allow the authors to study optimal allocation of cellular resources, including glycogen storage.

Dynamic FBA has also gained particular attention in the study of microbial interactions, and has this has led to the development of several dedicated software packages (Gomez et al. 2014, Zomorodi et al. 2014, Zhuang et al. 2011a), where some also takes into account spatial dynamics, such as COMETS (Harcombe et al. 2014, Dukovski et al. 2020) and BacArena (Bauer et al. 2017). COMETS is a powerful tool that now has been expanded to cover extracellular enzymes, evolution, demographic and growth noise, dynamic medium conditions and light absorption (Dukovski et al. 2020). The two latter features were developed in in Paper 5 to facilitate our dynamic FBA simulations of *Prochlorococcus*. The light availability is modelled using a sinusoidal wave form during the day (12 hours) and no light during the night time, while the actual light absorption is calculated using the Beer-Lambert law. Still, the simulations rely on a few simplifications: energy dissipation is not taken into account, and we assumed monochromatic light of 680 nm, and we did not weight the absorption light of the different photosystems according to their different absorption spectra. Also, we did not take into account that *Prochlorococcus* may be able to regulate the abundance of photosynthetic pigments to adjust light absorption. Despite these simplifications, the developed framework is a considerable contribution that will increase COMETS' relevance in future studies of phototrophic or mixotrophic organisms. As we show in Paper 5, the ability to simulate day-night cycles with dynamic allocation of carbon reveals aspects of *Prochlorococcus* beyond the scope of standard FBA.

COMETS is a population-based method where the growth of each organism in each cell of the spatial grid is considered a continuous increase in biomass. Thus, this made it difficult to take into account the diel life cycle of *Prochlorococcus* where cell division is mainly performed in the afternoon and early evening (Vaulot et al. 1995). This kind of cellular processes might have been easier to account for in an agent based framework such as BacAreana (Bauer et al. 2017) which runs an individual FBA model for each individual.

Chapter 6

Biosynthetic gene clusters

Biosynthetic gene clusters (BGCs) are groups of genes that are physically co-located on a genome and in total encode the enzymes required for the biosynthesis of a natural product (Medema et al. 2015). In addition to enzyme encoding genes, BGCs often also contain pathway-specific regulatory (Liu et al. 2013, Rigali et al. 2018), transport genes (Crits-Christoph et al. 2020), and genes responsible for resistance mechanisms protecting the cell or bacterium from self-damage caused by the final product (Alanjary et al. 2017). BGCs are most often found in bacteria or fungi, and the associated compounds comprise a wide range of different classes of natural products (Figure 6.1). These classes are of utmost interest in drug discovery because of their already demonstrated value as antibiotic, anticancer and antifungal drugs (Cragg and Newman 2013, Harvey 2008). In this context, the *Streptomyces* genus in the phylum of Actinobacteria is of particular interest as the most dominant source (Berdy 2005, Cimermancic et al. 2014). *Streptomyces coelicolor* is a model species for this phylum with well-established experimental protocols for reproducible cultivations and sampling of omics data, partly developed in previous SINTEF projects (Wentzel et al. 2012, Nieselt et al. 2010, Thomas et al. 2012).

As the developed pipeline in Paper 3 is currently limited to polyketides and non-ribosomal peptides (NRPs), produced by polyketides synthases (PKSs) and non-ribosomal peptide synthetases (NRPSs), respectively, we here focus on these two biosynthetic classes. These two classes are the most abundant classes of experimentally verified BGCs according to the MIBiG database (Kautsar et al. 2020; Figure 6.1), but probably not the most abundant classes in nature (Cimermancic et al. 2014). However, they are of immense importance in drug discovery (Masschelein et al. 2017).

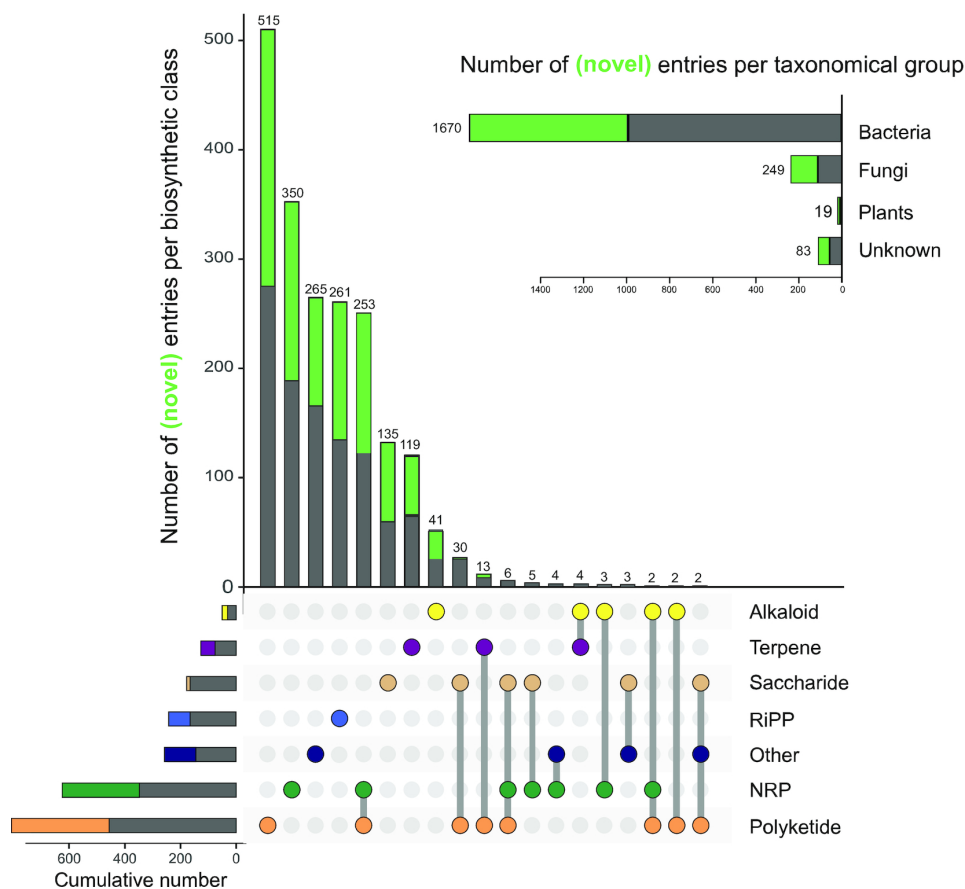


Figure 6.1: Distribution of the taxonomic kingdoms and biosynthetic classes of the BGCs in MIBiG. Content newly added in MIBiG 2.0 is shown in green. The connected circles in the lower panel indicate hybrid variants. Figure reused from [Kautsar et al. 2020](#) under the license [CC-BY-4.0](#).

Biosynthesis of NRPs and polyketides are conceptually similar, both governed by multidomain enzyme complexes that extend a growing polymer in a modular fashion (Figure 6.2, see [Challis and Naismith 2004](#) and [Keatinge-Clay 2012](#) for reviews of these two subjects, respectively). The general structure of an NRPS or PKS assembly line starts with a loading domain that attaches the first building block to a peptidyl carrier protein (PCP) or acyl carrier protein (ACP), respectively. Then, a number of elongation modules follow, each module containing at least three functional domains (abbreviations in the following text correspond to symbols in Figure 6.2). In NRPs these three domains are an adenylation domain (A) that activates and attaches a specific amino acid onto the the carrier domain

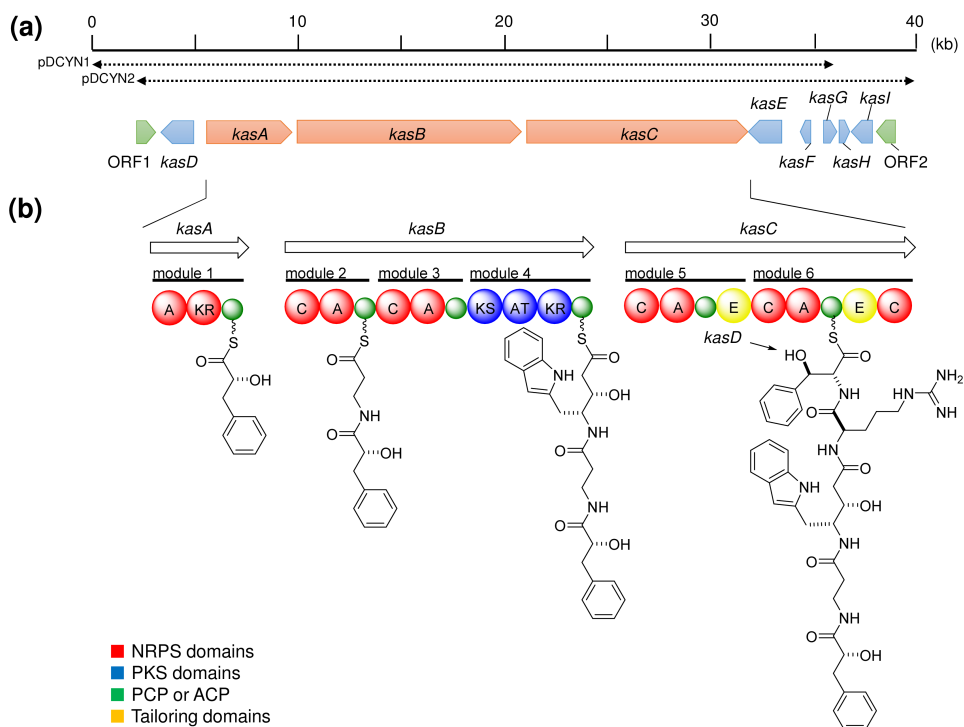


Figure 6.2: Illustrating the biosynthesis of a NRP-PSK hybrid compound, containing both NRPS and PKS modules. The top arrows represents genes, and each gene can contain several modules. Each module adds append one monomer to the growing polymer. The functional domains in each module are represented by coloured spheres. Figure reused from Nakashima et al. 2016 under the license CC-BY-4.0.

(PCP), and finally a condensation domain (C) that catalyzes the formation of peptide bonds to elongate the peptide chain. In PKSs the corresponding operations are carried out by an acyltransferase (AT) that attaches the acyl-group from an acyl-CoA onto the carrier domain (ACP), and finally a ketosynthase domain (KS) that catalyzes a Claisen condensation reaction to elongate the peptide chain. The final module contains a thioesterase or thioester reductase domain that cleaves the polymer from the carrier protein to release the product from the enzyme complex. Both NRPS and PKS modules can contain additional functional modules that modify the structure of the polyketide chain, and we refer to Table 1 in Paper 2 for a full overview of these domains, including their functions and abbreviations. NRPS-PKS hybrid BGCs are also frequent, combining functional modules from each biosynthetic class (Figure 6.2). This adds to the already huge diversity of natural products associated with these two classes of BGCs. Note that a particular type of PKSs, mostly found in fungi, uses the polyketide-elongating modules

iteratively (Cox et al. 2018), and differs from the collinear biosynthesis presented here.

The ability to identify potentially novel natural products from genome-sequences marks a conceptual change in drug discovery, formerly based on large-scale bioactivity screening. The diversity of computational approaches in this context is reviewed elsewhere (Medema and Fischbach 2015, Weber and Kim 2016). AntiSMASH is one of the more popular tools, defined as a high-confidence/low-novelty approach that uses DNA sequence signatures associated with known BGC domains to search for similar patterns in a target genome (Medema et al. 2011). For each identified BGC, antiSMASH details the genes present including functional modules and domains, and as such provides most of the information required to assemble the encoded biosynthetic pathway. In Paper 2 we use output from antiSMASH to automatically reconstruct PKS and NRPS biosynthesis pathways, thereby making antiSMASH a more accessible tool for *in silico* guided strain design of heterologous expression hosts.

Chapter 7

Summary of papers

This doctoral thesis contains one review and four research papers. Collectively, the four research papers present a diverse range of GEM applications, covering *in silico* strain-engineering, integration of omics data, dynamic simulations, and both biased and unbiased analyses of the steady-state solution space of the stoichiometric matrix. Because the review paper adds to the already presented background knowledge on GEM reconstruction and analysis, and thus sets the stage for Paper 2-5, this work is presented first. Paper 2, 3 and 4 are associated with the IN-BioPharm project, with an overarching goal of developing tools and knowledge required to ultimately improve on existing *S. coelicolor* strains for heterologous expression of BGCs. In Paper 5 the goal is to increase our general understating of how carbon allocation and release is determined by the diel cycle and changes in environmental parameters.

Paper 1 - Addressing Uncertainty in Genome-Scale Metabolic Model Reconstruction and Analysis

During the last two decades GEMs have proven to be versatile tools that are applicable to a wide range of research questions across different scientific disciplines (Gu et al. 2019). However, interpretation of model predictions are frequently hampered by the uncertainties associated with model reconstruction and analysis. In this review we present a comprehensive overview of the sources of uncertainty, and claim that the COBRA framework could benefit from a better representation of uncertainty, comprising both model reconstruction and analysis. We also describe existing approaches for representing or reducing model uncertainties, and suggest that ensemble-based or Bayesian approaches are potential frameworks that can be valuable in this context.

Paper 2 - Predicting Strain Engineering Strategies Using iKS1317: A Genome-Scale Metabolic Model of *Streptomyces coelicolor*

This is the first of our two papers concerning the reconstruction of a genome-scale metabolic model of *Streptomyces coelicolor*. The developed model, iKS1317, is based on content from two existing *S. coelicolor* GEMs (Kim et al. 2014, Alam et al. 2010), but we further expand on this knowledge by adding reactions from reaction databases, from the literature and from comparative genome alignment of genes with incomplete functional annotation. We assemble a set of growth and knockout phenotype data from the literature that we use to evaluate model performance and identify inaccuracies. Finally, we use two existing algorithms (Burgard et al. 2003, Lun et al. 2009) to predict strain-engineering strategies for increased availability of acetyl-CoA, an important precursor for polyketide biosynthesis.

Paper 3 - Automatic reconstruction of metabolic pathways from identified biosynthetic gene clusters

One of the main benefits of using GEMs to suggest strain-engineering strategies is the ability to account for the complete metabolic requirements for the synthesis of a particular compound, including cofactor and energy demands. However, one cannot fully leverage this framework in the context of heterologous expression of BGCs, unless one can make fairly accurate reconstructions of the associated biosynthetic pathways. Most of the information required for this pathway reconstruction is contained in the output of genome mining tools such as antiSMASH (Medema et al. 2011), but the manual effort and knowledge required to convert this data into a metabolic pathway are hampering the use of GEMs in this context. In this paper we develop a pipeline that automates this process, and we apply the pipeline to 943 metabolic pathways from a corresponding number of BGCs from the MIBiG database (Kautsar et al. 2020).

Paper 4 - Enzyme-Constrained Models and Omics Analysis of *Streptomyces coelicolor* Reveal Metabolic Changes that Enhance Heterologous Production

The first part of this paper covers the integration of iKS1317, presented in Paper 2, with two other updates of the *S. coelicolor* GEMs also published in 2018. Furthermore, we curate transport reactions, we inform the reversibility of model reactions based on calculated values for the change in Gibbs Free Energy, and we update the biomass equation with respect to prosthetic groups. GEM curation and development is a continuous process, and to facilitate further development and contributions from the scientific community we host the consensus *S. coelicolor* GEM publicly on GitHub. The remaining parts of the paper provide a comprehensive comparison of *S. coelicolor* M1152, a strain engineered for the purpose of

heterologous expression (Gomez-Escribano and Bibb 2011), with its wild-type (although devoid of plasmids) ancestor M145. We compare these two strains on the transcriptional level, on cultivation data, on proteome abundances and on the level of metabolic fluxes. The flux data are simulated by randomly sampling time- and strain-specific versions of the consensus *S. coelicolor* GEM, each model version with reaction fluxes constrained according to the measured enzyme abundances. We try to connect the genetic modifications of M1152 to the observed phenotype, but this is partly hampered by the complexity of regulatory mechanisms in *S. coelicolor* and the possible epistatic interactions. However, the results indicate that regulatory mechanisms are more important than the available precursor pool when it comes to heterologous expression and production of novel natural products. Furthermore, we describe several interesting discrepancies between the two strains that can serve as hypothesis in future research on this subject. For example, we suggest that the reduced growth rate in M1152 can derive from oxidative stress.

Paper 5 - Dynamic allocation of carbon storage and nutrient-dependent exudation in a revised genome-scale model of *Prochlorococcus*

The main goal of this paper is to understand how nutrients and availability of light affect the allocation and release of carbon in *Prochlorococcus*. First, we update the previous *Prochlorococcus* GEM (Casey et al. 2016) using a novel method that iteratively adds new reactions while ensuring complete connectivity of the network. Then, to understand how different nutrient environments affect the metabolism of *Prochlorococcus* we sample 10,000 nutrient environments that differ in availability of light, bicarbonate, phosphate and ammonium. For each environment we characterize the metabolic phenotype by using FBA, pFBA and FVA, and explore correlations between the four sampled "nutrients" and the predicted exudation of different metabolites. We find that glycogen storage or exudation of organic acids are favourable when the growth is limited by ammonium, while exudation of amino acids becomes more likely when the availability of phosphate is low. Finally, we run dFBA simulations through day-night cycles to find out how the diel availability of photons and dynamic storage of intracellular carbon in the form of glycogen affect its metabolism, both in terms of intracellular rewiring and with respect to the release of metabolites into the environment. In agreement with previous observations Zavřel et al. (2019), we find that glycogen is stored when growth is nutrient-limited, and we find that the consumption of intracellular glycogen is accompanied by an onset of the pentose phosphate pathway and exudation of organic acids.

Chapter 8

Conclusion

This work has covered a diverse use of genome-scale models, both in terms of different organisms and different applications. Initially, we focused on curation and evaluation of the iKS1317 GEM of *S. coelicolor* that ultimately evolved into Paper 2. Two other research groups published their own update of the *S. coelicolor* GEM within the same year. This proved the general interest in GEMs and constraint-based analyses from the *Streptomyces* scientific community, but it also revealed issues with independent efforts of individual research groups in model development. These three GEMs were based on the same template, but because of the different scope of each contribution, these three models now differed in content and to some extent also in namespace. To address the inconvenience of three different models of the same strain, we joined forces with one of the two other groups¹ in a completely open source and transparent development of a consensus *S. coelicolor* GEM, similar to previous initiatives on bakers' yeast (Lu et al. 2019) and Chinese Hamster Ovary (Hefzi et al. 2016). This consensus model is now hosted publicly on GitHub, and anyone can browse the complete history of the model development. Although this effort merged the content of the three preceding models and further improved the model with respect to transport reactions, reaction reversibility and prosthetic groups, there is still room for improvement of Sco-GEM. Known issues and potential improvements are tracked and discussed openly on the GitHub repository, and anyone can contribute to further development through pull requests by posting new issues. As such, this framework facilitate a continuous model development that extends beyond the time frames and capacity of individual projects and research groups. However, future value depends on active maintenance and involvement from the research community.

¹Eduard Kerkhoven, Chalmers University of Technology, Sweden

The scope of the INBioPharm project has defined the questions and applications of *S. coelicolor* GEMs within this PhD work. First, we applied two strain engineering algorithms, OptKnock (Burgard et al. 2003) and GDLS (Lun et al. 2009), to explore general strain engineering strategies towards increased production of polyketides. Similar to previous *S. coelicolor* host development (Gomez-Escribano and Bibb 2011), we wanted to search for strategies that would improve heterologous production in *S. coelicolor* in general, and not tailored to one specific compound. We opted at improving the availability of malonyl-CoA, the most used precursor for polyketide biosynthesis, through gene knockouts. Increased malonyl-CoA availability, through overexpression of the enzyme that converts acetyl-CoA to malonyl-CoA, had already been shown to increase the yield of the native polyketide actinorhodin (Ryu et al. 2006). Acetyl-CoA is the immediate source of malonyl-CoA. Therefore, we searched for single, double and triple reaction knockouts that would increase the rate through pyruvate dehydrogenase, forming acetyl-CoA from pyruvate and coenzyme A, across three different environments. We spanned different environments to increase the robustness of the predicted strain design strategies. Our results suggested double and triple knockout strategies that could provide a possible 2-fold increase in the flux through pyruvate dehydrogenase. However, none of the suggested knockout strategies resulted in a phenotype where the flux through pyruvate dehydrogenase was more strongly coupled to growth than in the wild-type, and in that sense the suggested strategies are not robust.

Although the suggested strategies have not been evaluated by actual genetic modifications and *in vitro* cultivations, the approach that was chosen has, in retrospect, a few caveats. First, increasing the flux through pyruvate dehydrogenase does not capture the fact that there is a real metabolite drain in the case where a final compound is actually produced and secreted. In our simulated experiments, the increased acetyl-CoA production has to be balanced by an increase in reactions that recycle and reuse this compound. Therefore, one may anticipate more biologically relevant predictions by optimization of a pseudo reaction that consumes the acyl moiety of acetyl- or malonyl-CoA, and releases coenzyme A back into the system, more closely replicating the incorporation of malonyl-CoA into the polyketide chain.

Furthermore, one of the benefits of using GEMs to guide strain development is the ability to account for more complex questions. Thus, as long as the metabolic pathway responsible for the synthesis of a particular compound is known, GEMs can easily be used to predict strain development strategies that is not limited to precursor pools, but also account for cofactor and energy demands. This motivated the development of the BigMeC pipeline for the automatic reconstruction of

metabolic pathways from identified biosynthetic gene clusters, presented in Paper 3. Leveraging this tool, we were able to explore single gene knockouts that increase metabolite production for 943 BGCs from the MIBiG database (Kautsar et al. 2020). Results indicate that single reaction knockouts are insufficient to provide a valuable increase in production rates in *S. coelicolor* for polyketides and non-ribosomal peptides. However, we foresee that this tool can be of value in future drug discovery, either in the initial selection of BGCs for further research, to suggest strain-engineering strategies, or to select the optimal BGC expression host.

Our second application of the *S. coelicolor* GEM aimed at unravelling the metabolic differences between the M145 strain (wild-type devoid of the two plasmids) and the derived M1152 strain that has been engineered towards increased heterologous production by a point mutation in the *rpoB* and removal of four biosynthetic gene clusters. To this end we used a recently developed framework to incorporate enzymatic parameters (catalytic rate and enzyme mass), time-series proteome and cultivation data for both strains. The data was collected at regular intervals during a batch fermentation covering the initial lag and growth phase, depletion of phosphate and the onset of secondary metabolism. The measured protein levels at each time step for each strain were used along with the cultivation data to constrain the solution space. Because the cellular objective is likely to change throughout the cultivations, we opted at an unbiased approach to describe how these constraints shaped the metabolism both between the two strain and throughout the cultivations. We contrasted the model predictions with transcriptome data to get a more holistic understanding. With this approach we were able to capture both known and novel strain differences. These observations will be important in future development of *S. coelicolor* for heterologous expression of BGCs. For example, the absence of Actinorhodin may affect the redox-regulator SoxR, and indeed, the omics data indicate that *S. coelicolor* M1152 suffers from oxidative stress. This may explain the slowed growth of this strain. Furthermore, it is possible that the deletion of *ScbR2*, which is a part of the *cpk* BGC responsible for the production of Coelimycin P1, affects global regulators. Correspondingly, global differences in the proteome and transcriptome of these two strain indicate that global regulators are affected by the genetic alterations in M1152. However, because of the large genetic difference between these two strains it is difficult to pinpoint whether the observed differences are caused by the *rpoB* mutation or by absence of the four BGCs. The extremely complex and not fully described regulatory mechanisms in *S. coelicolor* adds to this difficulty. In retrospect, it is clear that it would have been beneficial to include the intermediate strain M1146. Nevertheless, we suggest that regulatory mechanisms may play a more important role than the availability of precursor pools for increased production of bioactive molecules encoded by BGCs

heterologously expressed in *S. coelicolor*.

The phototrophic *Prochlorococcus* encourages different research questions compared to *S. coelicolor*, which mostly have been directed towards drug discovery. As a primary producer, it is of utmost interest to understand how various nutritional access and environmental conditions affect which, and to what extent, different organic compounds are released into the ocean space. We targeted this question by randomly sampling different environments, with different access to sunlight, ammonium, phosphate and bicarbonate. To replicate the unavoidable inflow of photons and bicarbonate we used a variant of the standard FBA where we modified the bounds in the corresponding exchange reactions to force a certain uptake. To make the results interpretable we used unsupervised machine learning to cluster and project the data points onto the 2D space. With a recent clustering algorithm (McInnes et al. 2017) we could identify six main clusters that we interpreted as typical phenotypes. This analysis, however, could not reveal aspects related to the diurnal periodicity of *Prochlorococcus*' metabolism that is dictated by the availability of photons from sunlight. To capture this phenomenon, we both rewired the representation of glycogen metabolism in the *Prochlorococcus* GEM iSO595 to allow carbon storage and depletion, and we incorporated periodic environmental conditions and light absorption into COMETS (Dukovski et al. 2020, Harcombe et al. 2014). These simulations provide novel hypothesis for how the metabolic fluxes are reorganized in *Prochlorococcus* during the shift from photosynthesis and accumulation of glycogen during the day to depletion of the accumulated glycogen during the afternoon and night. A qualitative comparison with growth data of a closely related organism indicate that our simulations are in the right ballpark, and that this framework will be useful in future research of phototrophic organisms. One open question, that we now can, and should address is what metabolic interactions that governs the mutual benefit in co-cultivations of *Prochlorococcus* and heterotrophs (Roth-Rosenberg et al. 2020, Sher et al. 2011).

Although the work on alveolar macrophages is yet to be concluded I briefly summarize the current conclusions here. First, unsteady-FBA seems to require fairly large changes in intracellular metabolites levels over a short time span to significantly constrain and, thereby improve, FBA predictions. The 2 hour resolution and intracellular changes observed in the stimulated human alveolar macrophages did not suffice. Therefore, TFA seems like a more appropriate method, and preliminary results show that the different metabolite levels actually have an impact the calculated changes in Gibbs Free Energy and thus the likely direction of intracellular. Another difficulty of this study is the lack of growth: the stimulated macrophages perform cellular maintenance, but they don't grow. We have currently modelled this as simply a very low growth rate, similar to previous work (Bordbar et al.

2017), but it is not clear whether or not this is appropriate. Anyhow, it will be exciting to dig further into these questions, and ultimately elucidate how stimulation with different ligands affect the metabolism of these cells on a global level.

The diversity of GEM applications encountered during these last four years, combined with the exposure to several research groups, have displayed the strengths and weaknesses of the COBRA field. These strengths and weaknesses are not readily grasped by scientists outside the field, and it is my impression that inadequate presentations of the scope and limitations of this field can result in distorted expectations and challenging interdisciplinary collaborations. For example, it may seem contradicting that metabolomics are not directly comparable with the output from metabolic models. However, these interdisciplinary collaborations between experimental and computational biologists are extremely valuable, in particular for more applied research and with respect to the increased scope of GEMs towards non-model species and microbial consortia. Successful interdisciplinary collaborations require a minimum understanding of the fundamental limitations and assumptions associated with each methodology. Therefore, rather than downplaying the assumptions and uncertainties associated with GEM predictions, these aspects should be embraced and clearly represented. In Paper 1 we present the current state of the COBRA field in terms of sources of uncertainty and approaches that address this topic. This review is not meant to discredit the COBRA field. Rather, we review this topic to facilitate further development of frameworks that can consolidate the trust in GEM research.

Chapter 9

Outlook

The COBRA field has seen a massive development in scope during the last decades. One major reason for this success is, in my opinion, the ability of GEMs to strike a good balance between usability and complexity. Although the framework has clear limitations, in particular the lack of kinetic information and metabolite concentrations, the literature demonstrate the applicability of this framework to a diverse range of medical, biological and ecological questions. These limitations are actually enabling the genome-scale of these metabolic models, in contrast to adjacent fields such as kinetic and whole cell modelling which both are weighed down by many, and often unknown, parameters (Saa and Nielsen 2017, Babbie and Stumpf 2017).

The scope of the COBRA framework has recently been expanded with the development of ME-models, which expand on the standard GEM formulation to account for transcription and translation (Lloyd et al. 2018, Ebrahim et al. 2013). With these models one can address new questions, e.g. how does reactive oxygen species or acid stress affect the metabolism of *E. coli* (Du et al. 2019, Yang et al. 2019), but ME-models are not only computationally more costly, but they also rely on parameters that are not readily available for non-model organisms. There is actually in general a challenge in the COBRA field to develop and apply models to study non-model species, not only because of the lack of species-specific knowledge, but also because of the significant amount of curation required to make high-quality models. However, the latter concern is being addressed by a constant development of novel model reconstruction softwares (Mendoza et al. 2019). As we mention in Paper 1, we believe that these reconstruction softwares can improve from a better representation of uncertainties, e.g. as an ensemble of models as in Medusa (Medlock et al. 2020). The first concern might be addressed by improved

tools and methods for assessing growth or knockout phenotypes for multiple species across a wide range of environments (van Leeuwen et al. 2020, Kehe et al. 2019), that can be incorporated in novel automatic reconstruction pipelines.

With the rapid improvement and reduced cost of sequencing technology, it is likely that omics integration and interpretation will become an even more important application of GEMs. Recent methodological advances use large sets of transcriptome data to infer transcriptional regulatory networks (Sastry et al. 2019). Complete regulatory networks and GEM integration might be crucial to allow accurate predictions for organisms such as *S. coelicolor* which features a complex life cycle and regulatory mechanisms that convolute genotype-phenotype associations. Another interesting breakthrough is the impressive accuracy of AlphaFold 2 in predicting protein folding from the amino acid sequence¹. Protein structures can expand the scope of GEM applications (Brunk et al. 2016), and aid in the prediction of kinetic coefficients (Heckmann et al. 2018) which are crucial for an extended use of enzyme-constrained GEMs for none-model species. In total, these recent developments create exciting opportunities for further use of GEMs and development of the COBRA framework.

¹<https://doi.org/10.1038/d41586-020-03348-4>

Chapter 10

Bibliography

- Adadi, R., Volkmer, B., Milo, R., Heinemann, M., and Shlomi, T. (2012). Prediction of microbial growth rate versus biomass yield by a metabolic network with kinetic parameters. *PLoS Comput Biol*, 8(7):e1002575.
- Agren, R., Bordel, S., Mardinoglu, A., Pornputtapong, N., Nookaew, I., and Nielsen, J. (2012). Reconstruction of genome-scale active metabolic networks for 69 human cell types and 16 cancer types using init. *PLoS Comput Biol*, 8(5):e1002518.
- Agren, R., Mardinoglu, A., Asplund, A., Kampf, C., Uhlen, M., and Nielsen, J. (2014). Identification of anticancer drugs for hepatocellular carcinoma through personalized genome-scale metabolic modeling. *Molecular systems biology*, 10(3):721.
- Åkesson, M., Förster, J., and Nielsen, J. (2004). Integration of gene expression data into genome-scale metabolic models. *Metabolic engineering*, 6(4):285–293.
- Alam, M. T., Merlo, M. E., Hodgson, D. A., Wellington, E. M., Takano, E., Breitling, R., Consortium, S., et al. (2010). Metabolic modeling and analysis of the metabolic switch in streptomyces coelicolor. *BMC genomics*, 11(1):202.
- Alanjary, M., Kronmiller, B., Adamek, M., Blin, K., Weber, T., Huson, D., Philmus, B., and Ziemert, N. (2017). The antibiotic resistant target seeker (arts), an exploration engine for antibiotic cluster prioritization and novel drug target discovery. *Nucleic acids research*, 45(W1):W42–W48.
- Almaas, E., Kovacs, B., Vicsek, T., Oltvai, Z. N., and Barabási, A.-L. (2004). Global organization of metabolic fluxes in the bacterium escherichia coli. *Nature*, 427(6977):839–843.

- Aurich, M. K., Fleming, R. M., and Thiele, I. (2016). Metabotools: a comprehensive toolbox for analysis of genome-scale metabolic models. *Frontiers in physiology*, 7:327.
- Aziz, R. K., Bartels, D., Best, A. A., DeJongh, M., Disz, T., Edwards, R. A., Formsma, K., Gerdes, S., Glass, E. M., Kubal, M., et al. (2008). The rast server: rapid annotations using subsystems technology. *BMC genomics*, 9(1):1–15.
- Babtie, A. C. and Stumpf, M. P. (2017). How to deal with parameters for whole-cell modelling. *Journal of The Royal Society Interface*, 14(133):20170237.
- Bar-Even, A., Noor, E., Savir, Y., Liebermeister, W., Davidi, D., Tawfik, D. S., and Milo, R. (2011). The moderately efficient enzyme: evolutionary and physico-chemical trends shaping enzyme parameters. *Biochemistry*, 50(21):4402–4410.
- Baroukh, C., Muñoz-Tamayo, R., Bernard, O., and Steyer, J.-P. (2015). Mathematical modeling of unicellular microalgae and cyanobacteria metabolism for biofuel production. *Current opinion in biotechnology*, 33:198–205.
- Baroukh, C., Muñoz-Tamayo, R., Steyer, J.-P., and Bernard, O. (2014). Drum: a new framework for metabolic modeling under non-balanced growth. application to the carbon metabolism of unicellular microalgae. *PLoS one*, 9(8):e104499.
- Basan, M., Hui, S., Okano, H., Zhang, Z., Shen, Y., Williamson, J. R., and Hwa, T. (2015). Overflow metabolism in escherichia coli results from efficient proteome allocation. *Nature*, 528(7580):99–104.
- Bauer, E., Zimmermann, J., Baldini, F., Thiele, I., and Kaleta, C. (2017). Bacarena: Individual-based metabolic modeling of heterogeneous microbes in complex communities. *PLoS computational biology*, 13(5):e1005544.
- Beck, A. E., Hunt, K. A., and Carlson, R. P. (2018). Measuring cellular biomass composition for computational biology applications. *Processes*, 6(5):38.
- Beg, Q. K., Vazquez, A., Ernst, J., de Menezes, M. A., Bar-Joseph, Z., Barabási, A.-L., and Oltvai, Z. N. (2007). Intracellular crowding defines the mode and sequence of substrate uptake by escherichia coli and constrains its metabolic activity. *Proceedings of the National Academy of Sciences*, 104(31):12663–12668.
- Bentley, S. D., Chater, K. F., Cerdeño-Tárraga, A.-M., Challis, G. L., Thomson, N., James, K. D., Harris, D. E., Quail, M. A., Kieser, H., Harper, D., et al. (2002). Complete genome sequence of the model actinomycete streptomyces coelicolor a3 (2). *Nature*, 417(6885):141–147.

-
- Berdy, J. (2005). Bioactive microbial metabolites. *The Journal of antibiotics*, 58(1):1–26.
- Biggs, M. B. and Papin, J. A. (2017). Managing uncertainty in metabolic network structure and improving predictions using ensemblefba. *PLoS computational biology*, 13(3):e1005413.
- Blattner, F. R., Plunkett, G., Bloch, C. A., Perna, N. T., Burland, V., Riley, M., Collado-Vides, J., Glasner, J. D., Rode, C. K., Mayhew, G. F., et al. (1997). The complete genome sequence of escherichia coli k-12. *science*, 277(5331):1453–1462.
- Bordbar, A., Yurkovich, J. T., Paglia, G., Rolfsson, O., Sigurjónsson, Ó. E., and Palsson, B. O. (2017). Elucidating dynamic metabolic physiology through network integration of quantitative time-course metabolomics. *Scientific reports*, 7:46249.
- Bordel, S., Agren, R., and Nielsen, J. (2010). Sampling the solution space in genome-scale metabolic networks reveals transcriptional regulation in key enzymes. *PLoS Comput Biol*, 6(7):e1000859.
- Brunk, E., Mih, N., Monk, J., Zhang, Z., O'Brien, E. J., Bliven, S. E., Chen, K., Chang, R. L., Bourne, P. E., and Palsson, B. O. (2016). Systems biology of the structural proteome. *BMC systems biology*, 10(1):1–16.
- Buchholz, A., Hurlbauss, J., Wandrey, C., and Takors, R. (2002). Metabolomics: quantification of intracellular metabolite dynamics. *Biomolecular engineering*, 19(1):5–15.
- Burgard, A. P. and Maranas, C. D. (2001). Probing the performance limits of the escherichia coli metabolic network subject to gene additions or deletions. *Biotechnology and bioengineering*, 74(5):364–375.
- Burgard, A. P., Pharkya, P., and Maranas, C. D. (2003). Optknock: a bilevel programming framework for identifying gene knockout strategies for microbial strain optimization. *Biotechnology and bioengineering*, 84(6):647–657.
- Cantarel, B. L., Korf, I., Robb, S. M., Parra, G., Ross, E., Moore, B., Holt, C., Alvarado, A. S., and Yandell, M. (2008). Maker: an easy-to-use annotation pipeline designed for emerging model organism genomes. *Genome research*, 18(1):188–196.
- Carey, M. A., Dräger, A., Beber, M. E., Papin, J. A., and Yurkovich, J. T. (2020). Community standards to facilitate development and address challenges in metabolic modeling. *Molecular Systems Biology*, 16(8):e9235.

- Casey, J. R., Mardinoglu, A., Nielsen, J., and Karl, D. M. (2016). Adaptive evolution of phosphorus metabolism in prochlorococcus. *Msystems*, 1(6).
- Challis, G. L. and Naismith, J. H. (2004). Structural aspects of non-ribosomal peptide biosynthesis. *Current opinion in structural biology*, 14(6):748–756.
- Cho, J. S., Gu, C., Han, T. H., Ryu, J. Y., and Lee, S. Y. (2019). Reconstruction of context-specific genome-scale metabolic models using multiomics data to study metabolic rewiring. *Current Opinion in Systems Biology*, 15:1–11.
- Cimermancic, P., Medema, M. H., Claesen, J., Kurita, K., Brown, L. C. W., Mavrommatis, K., Pati, A., Godfrey, P. A., Koehrsen, M., Clardy, J., et al. (2014). Insights into secondary metabolism from a global analysis of prokaryotic biosynthetic gene clusters. *Cell*, 158(2):412–421.
- Colijn, C., Brandes, A., Zucker, J., Lun, D. S., Weiner, B., Farhat, M. R., Cheng, T.-Y., Moody, D. B., Murray, M., and Galagan, J. E. (2009). Interpreting expression data with metabolic flux models: predicting mycobacterium tuberculosis mycolic acid production. *PLoS Comput Biol*, 5(8):e1000489.
- Collins, F. S., Morgan, M., and Patrinos, A. (2003). The human genome project: lessons from large-scale biology. *Science*, 300(5617):286–290.
- Collins, S. B., Reznik, E., and Segrè, D. (2012). Temporal expression-based analysis of metabolism. *PLoS Comput Biol*, 8(11):e1002781.
- Cox, R. J., Skellam, E., and Williams, K. (2018). Biosynthesis of fungal polyketides. In *Physiology and Genetics*, pages 385–412. Springer.
- Cragg, G. M. and Newman, D. J. (2013). Natural products: a continuing source of novel drug leads. *Biochimica et Biophysica Acta (BBA)-General Subjects*, 1830(6):3670–3695.
- Crits-Christoph, A., Bhattacharya, N., Olm, M. R., Song, Y. S., and Banfield, J. F. (2020). Transporter genes in biosynthetic gene clusters predict metabolite characteristics and siderophore activity. *bioRxiv*.
- Cruz, F., Faria, J. P., Rocha, M., Rocha, I., and Dias, O. (2020). A review of methods for the reconstruction and analysis of integrated genome-scale models of metabolism and regulation. *Biochemical Society Transactions*, 48(5):1889–1903.
- Davidi, D., Noor, E., Liebermeister, W., Bar-Even, A., Flamholz, A., Tumbler, K., Barenholz, U., Goldenfeld, M., Shlomi, T., and Milo, R. (2016). Global

-
- characterization of in vivo enzyme catalytic rates and their correspondence to in vitro k_{cat} measurements. *Proceedings of the National Academy of Sciences*, 113(12):3401–3406.
- De Martino, D., Mori, M., and Parisi, V. (2015). Uniform sampling of steady states in metabolic networks: heterogeneous scales and rounding. *PloS one*, 10(4):e0122670.
- Du, B., Yang, L., Lloyd, C. J., Fang, X., and Palsson, B. O. (2019). Genome-scale model of metabolism and gene expression provides a multi-scale description of acid stress responses in escherichia coli. *PLoS computational biology*, 15(12):e1007525.
- Dukovski, I., Bajić, D., Chacón, J. M., Quintin, M., Vila, J. C., Sulheim, S., Pacheco, A. R., Bernstein, D. B., Rieh, W. J., Korolev, K. S., et al. (2020). Computation of microbial ecosystems in time and space (comets): An open source collaborative platform for modeling ecosystems metabolism. *arXiv pre-print arXiv:2009.01734*.
- Ebrahim, A., Lerman, J. A., Palsson, B. O., and Hyduke, D. R. (2013). Cobrapy: constraints-based reconstruction and analysis for python. *BMC systems biology*, 7(1):74.
- Edwards, J. and Palsson, B. (2000). The escherichia coli mg1655 in silico metabolic genotype: its definition, characteristics, and capabilities. *Proceedings of the National Academy of Sciences*, 97(10):5528–5533.
- Edwards, J. S., Covert, M., and Palsson, B. (2002). Metabolic modelling of microbes: the flux-balance approach. *Environmental microbiology*, 4(3):133–140.
- Edwards, J. S. and Palsson, B. O. (1999). Systems properties of the haemophilus influenzae metabolic genotype. *Journal of Biological Chemistry*, 274(25):17410–17416.
- Fang, X., Lloyd, C. J., and Palsson, B. O. (2020). Reconstructing organisms in silico: genome-scale models and their emerging applications. *Nature Reviews Microbiology*, 18(12):731–743.
- Faria, J. P., Overbeek, R., Xia, F., Rocha, M., Rocha, I., and Henry, C. S. (2014). Genome-scale bacterial transcriptional regulatory networks: reconstruction and integrated analysis with metabolic models. *Briefings in bioinformatics*, 15(4):592–611.

- Feist, A. M., Zielinski, D. C., Orth, J. D., Schellenberger, J., Herrgard, M. J., and Palsson, B. Ø. (2010). Model-driven evaluation of the production potential for growth-coupled products of *escherichia coli*. *Metabolic engineering*, 12(3):173–186.
- Fell, D. A. and Small, J. R. (1986). Fat synthesis in adipose tissue. an examination of stoichiometric constraints. *Biochemical Journal*, 238(3):781–786.
- Field, C. B., Behrenfeld, M. J., Randerson, J. T., and Falkowski, P. (1998). Primary production of the biosphere: integrating terrestrial and oceanic components. *science*, 281(5374):237–240.
- Fleischmann, R. D., Adams, M. D., White, O., Clayton, R. A., Kirkness, E. F., Kerlavage, A. R., Bult, C. J., Tomb, J.-F., Dougherty, B. A., Merrick, J. M., et al. (1995). Whole-genome random sequencing and assembly of *haemophilus influenzae* rd. *Science*, 269(5223):496–512.
- Flombaum, P., Gallegos, J. L., Gordillo, R. A., Rincón, J., Zabala, L. L., Jiao, N., Karl, D. M., Li, W. K., Lomas, M. W., Veneziano, D., et al. (2013). Present and future global distributions of the marine cyanobacteria *prochlorococcus* and *synechococcus*. *Proceedings of the National Academy of Sciences*, 110(24):9824–9829.
- Förster, J., Famili, I., Palsson, B. Ø., and Nielsen, J. (2003). Large-scale evaluation of in silico gene deletions in *saccharomyces cerevisiae*. *OMICS A Journal of Integrative Biology*, 7(2):193–202.
- Gehlenborg, N., O’donoghue, S. I., Baliga, N. S., Goesmann, A., Hibbs, M. A., Kitano, H., Kohlbacher, O., Neuweger, H., Schneider, R., Tenenbaum, D., et al. (2010). Visualization of omics data for systems biology. *Nature methods*, 7(3):S56–S68.
- Gianchandani, E. P., Oberhardt, M. A., Burgard, A. P., Maranas, C. D., and Papin, J. A. (2008). Predicting biological system objectives de novo from internal state measurements. *BMC bioinformatics*, 9(1):43.
- Goelzer, A., Fromion, V., and Scorletti, G. (2011). Cell design in bacteria as a convex optimization problem. *Automatica*, 47(6):1210–1218.
- Gomez, J. A., Höffner, K., and Barton, P. I. (2014). Dfbalab: a fast and reliable matlab code for dynamic flux balance analysis. *BMC bioinformatics*, 15(1):409.
- Gomez-Escribano, J. P. and Bibb, M. J. (2011). Engineering *streptomyces coelicolor* for heterologous expression of secondary metabolite gene clusters. *Microbial Biotechnology*, 4(2):207–215.

-
- Gu, C., Kim, G. B., Kim, W. J., Kim, H. U., and Lee, S. Y. (2019). Current status and applications of genome-scale metabolic models. *Genome biology*, 20(1):121.
- Gudmundsson, S. and Thiele, I. (2010). Computationally efficient flux variability analysis. *BMC bioinformatics*, 11(1):489.
- Hadadi, N., Pandey, V., Chiappino-Pepe, A., Morales, M., Gallart-Ayala, H., Mehl, F., Ivanisevic, J., Sentchilo, V., and van der Meer, J. R. (2020). Mechanistic insights into bacterial metabolic reprogramming from omics-integrated genome-scale models. *NPJ systems biology and applications*, 6(1):1–11.
- Haraldsdóttir, H. S., Cousins, B., Thiele, I., Fleming, R. M., and Vempala, S. (2017). Chrr: coordinate hit-and-run with rounding for uniform sampling of constraint-based models. *Bioinformatics*, 33(11):1741–1743.
- Harcombe, W. R., Riehl, W. J., Dukovski, I., Granger, B. R., Betts, A., Lang, A. H., Bonilla, G., Kar, A., Leiby, N., Mehta, P., et al. (2014). Metabolic resource allocation in individual microbes determines ecosystem interactions and spatial dynamics. *Cell reports*, 7(4):1104–1115.
- Harvey, A. L. (2008). Natural products in drug discovery. *Drug discovery today*, 13(19-20):894–901.
- 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. (2018). Machine learning applied to enzyme turnover numbers reveals protein structural correlates and improves metabolic models. *Nature communications*, 9(1):1–10.
- Hefzi, H., Ang, K. S., Hanscho, M., Bordbar, A., Ruckerbauer, D., Lakshmanan, M., Orellana, C. A., Baycin-Hizal, D., Huang, Y., Ley, D., et al. (2016). A consensus genome-scale reconstruction of chinese hamster ovary cell metabolism. *Cell Systems*, 3(5):434–443.
- Henry, C. S., Broadbelt, L. J., and Hatzimanikatis, V. (2007). Thermodynamics-based metabolic flux analysis. *Biophysical journal*, 92(5):1792–1805.
- Ibarra, R. U., Edwards, J. S., and Palsson, B. O. (2002). *Escherichia coli* k-12 undergoes adaptive evolution to achieve in silico predicted optimal growth. *Nature*, 420(6912):186–189.
- Jankowski, M. D., Henry, C. S., Broadbelt, L. J., and Hatzimanikatis, V. (2008). Group contribution method for thermodynamic analysis of complex metabolic networks. *Biophysical journal*, 95(3):1487–1499.

- Jayapal, K. P., Philp, R. J., Kok, Y.-J., Yap, M. G., Sherman, D. H., Griffin, T. J., and Hu, W.-S. (2008). Uncovering genes with divergent mrna-protein dynamics in streptomyces coelicolor. *PLoS one*, 3(5):e2097.
- Jensen, P. A. and Papin, J. A. (2011). Functional integration of a metabolic network model and expression data without arbitrary thresholding. *Bioinformatics*, 27(4):541–547.
- Kanehisa, M. and Goto, S. (2000). Kegg: kyoto encyclopedia of genes and genomes. *Nucleic acids research*, 28(1):27–30.
- Karp, P. D., Riley, M., Paley, S. M., and Pellegrini-Toole, A. (2002). The metacyc database. *Nucleic acids research*, 30(1):59–61.
- Kautsar, S. A., Blin, K., Shaw, S., Navarro-Muñoz, J. C., Terlouw, B. R., van der Hooft, J. J., Van Santen, J. A., Tracanna, V., Suarez Duran, H. G., Pascal Andreu, V., et al. (2020). Mibig 2.0: a repository for biosynthetic gene clusters of known function. *Nucleic acids research*, 48(D1):D454–D458.
- Keasling, J. D. (2010). Manufacturing molecules through metabolic engineering. *Science*, 330(6009):1355–1358.
- Keatinge-Clay, A. T. (2012). The structures of type i polyketide synthases. *Natural product reports*, 29(10):1050–1073.
- Kehe, J., Kulesa, A., Ortiz, A., Ackerman, C. M., Thakku, S. G., Sellers, D., Kuehn, S., Gore, J., Friedman, J., and Blainey, P. C. (2019). Massively parallel screening of synthetic microbial communities. *Proceedings of the National Academy of Sciences*, 116(26):12804–12809.
- Kim, M., Sang Yi, J., Kim, J., Kim, J.-N., Kim, M. W., and Kim, B.-G. (2014). Reconstruction of a high-quality metabolic model enables the identification of gene overexpression targets for enhanced antibiotic production in streptomyces coelicolor a3 (2). *Biotechnology journal*, 9(9):1185–1194.
- King, Z. A., Lu, J., Dräger, A., Miller, P., Federowicz, S., Lerman, J. A., Ebrahim, A., Palsson, B. O., and Lewis, N. E. (2016). Bigg models: A platform for integrating, standardizing and sharing genome-scale models. *Nucleic acids research*, 44(D1):D515–D522.
- Klamt, S., Regensburger, G., Gerstl, M. P., Jungreuthmayer, C., Schuster, S., Mahadevan, R., Zanghellini, J., and Müller, S. (2017). From elementary flux modes to elementary flux vectors: Metabolic pathway analysis with arbitrary linear flux constraints. *PLoS computational biology*, 13(4):e1005409.

-
- Knoop, H., Gründel, M., Zilliges, Y., Lehmann, R., Hoffmann, S., Lockau, W., and Steuer, R. (2013). Flux balance analysis of cyanobacterial metabolism: the metabolic network of *synechocystis* sp. pcc 6803. *PLoS Comput Biol*, 9(6):e1003081.
- Lachance, J.-C., Lloyd, C. J., Monk, J. M., Yang, L., Sastry, A. V., Seif, Y., Palsson, B. O., Rodrigue, S., Feist, A. M., King, Z. A., et al. (2019). Bofdat: Generating biomass objective functions for genome-scale metabolic models from experimental data. *PLoS computational biology*, 15(4):e1006971.
- Lee, D., Smallbone, K., Dunn, W. B., Murabito, E., Winder, C. L., Kell, D. B., Mendes, P., and Swainston, N. (2012). Improving metabolic flux predictions using absolute gene expression data. *BMC systems biology*, 6(1):73.
- Lewin, H. A., Robinson, G. E., Kress, W. J., Baker, W. J., Coddington, J., Crandall, K. A., Durbin, R., Edwards, S. V., Forest, F., Gilbert, M. T. P., Goldstein, M. M., Grigoriev, I. V., Hackett, K. J., Haussler, D., Jarvis, E. D., Johnson, W. E., Patrinos, A., Richards, S., Castilla-Rubio, J. C., van Sluys, M.-A., Soltis, P. S., Xu, X., Yang, H., and Zhang, G. (2018). Earth biogenome project: Sequencing life for the future of life. *Proceedings of the National Academy of Sciences*, 115(17):4325–4333.
- 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., et al. (2010). Omic data from evolved *e. coli* are consistent with computed optimal growth from genome-scale models. *Molecular systems biology*, 6(1):390.
- Lewis, N. E., Nagarajan, H., and Palsson, B. O. (2012). Constraining the metabolic genotype–phenotype relationship using a phylogeny of in silico methods. *Nature Reviews Microbiology*, 10(4):291–305.
- Lieven, C., Beber, M. E., Olivier, B. G., Bergmann, F. T., Ataman, M., Babaei, P., Bartell, J. A., Blank, L. M., Chauhan, S., Correia, K., et al. (2020). Memote for standardized genome-scale metabolic model testing. *Nature biotechnology*, 38(3):272–276.
- Liu, G., Chater, K. F., Chandra, G., Niu, G., and Tan, H. (2013). Molecular regulation of antibiotic biosynthesis in streptomyces. *Microbiology and molecular biology reviews*, 77(1):112–143.
- Lloyd, C. J., Ebrahim, A., Yang, L., King, Z. A., Catoiu, E., O’Brien, E. J., Liu, J. K., and Palsson, B. O. (2018). Cobrame: A computational framework for genome-scale models of metabolism and gene expression. *PLoS computational biology*, 14(7):e1006302.

- Lu, H., Li, F., Sánchez, B. J., Zhu, Z., Li, G., Domenzain, I., Marcišauskas, S., Anton, P. M., Lappa, D., Lieven, C., et al. (2019). A consensus *s. cerevisiae* metabolic model yeast8 and its ecosystem for comprehensively probing cellular metabolism. *Nature communications*, 10(1):1–13.
- Lun, D. S., Rockwell, G., Guido, N. J., Baym, M., Kelner, J. A., Berger, B., Galagan, J. E., and Church, G. M. (2009). Large-scale identification of genetic design strategies using local search. *molecular systems biology*, 5(1):296.
- Maarleveld, T. R., Wortel, M. T., Olivier, B. G., Teusink, B., and Bruggeman, F. J. (2015). Interplay between constraints, objectives, and optimality for genome-scale stoichiometric models. *PLoS Comput Biol*, 11(4):e1004166.
- Machado, D. and Herrgård, M. (2014). Systematic evaluation of methods for integration of transcriptomic data into constraint-based models of metabolism. *PLoS Comput Biol*, 10(4):e1003580.
- Mahadevan, R., Edwards, J. S., and Doyle III, F. J. (2002). Dynamic flux balance analysis of diauxic growth in *escherichia coli*. *Biophysical journal*, 83(3):1331–1340.
- Mahadevan, R. and Schilling, C. (2003). The effects of alternate optimal solutions in constraint-based genome-scale metabolic models. *Metabolic engineering*, 5(4):264–276.
- Maia, P., Rocha, M., and Rocha, I. (2016). In silico constraint-based strain optimization methods: the quest for optimal cell factories. *Microbiology and Molecular Biology Reviews*, 80(1):45–67.
- Maier, T., Schmidt, A., Güell, M., Kühner, S., Gavin, A.-C., Aebersold, R., and Serrano, L. (2011). Quantification of mrna and protein and integration with protein turnover in a bacterium. *Molecular systems biology*, 7(1):511.
- Masschelein, J., Jenner, M., and Challis, G. L. (2017). Antibiotics from gram-negative bacteria: a comprehensive overview and selected biosynthetic highlights. *Natural product reports*, 34(7):712–783.
- McInnes, L., Healy, J., and Astels, S. (2017). hdbscan: Hierarchical density based clustering. *Journal of Open Source Software*, 2(11):205.
- Medema, M. H., Blin, K., Cimermancic, P., de Jager, V., Zakrzewski, P., Fischbach, M. A., Weber, T., Takano, E., and Breitling, R. (2011). antismash: rapid identification, annotation and analysis of secondary metabolite biosynthesis gene clusters in bacterial and fungal genome sequences. *Nucleic acids research*, 39(suppl_2):W339–W346.

-
- Medema, M. H. and Fischbach, M. A. (2015). Computational approaches to natural product discovery. *Nature chemical biology*, 11(9):639.
- Medema, M. H., Kottmann, R., Yilmaz, P., Cummings, M., Biggins, J. B., Blin, K., De Bruijn, I., Chooi, Y. H., Claesen, J., Coates, R. C., et al. (2015). Minimum information about a biosynthetic gene cluster. *Nature chemical biology*, 11(9):625–631.
- Medlock, G. L., Moutinho, T. J., and Papin, J. A. (2020). Medusa: software to build and analyze ensembles of genome-scale metabolic network reconstructions. *PLoS computational biology*, 16(4):e1007847.
- Medlock, G. L. and Papin, J. A. (2020). Guiding the refinement of biochemical knowledgebases with ensembles of metabolic networks and machine learning. *Cell systems*, 10(1):109–119.
- Megchelenbrink, W., Huynen, M., and Marchiori, E. (2014). optgpsampler: an improved tool for uniformly sampling the solution-space of genome-scale metabolic networks. *PLoS one*, 9(2):e86587.
- Mendoza, S. N., Olivier, B. G., Molenaar, D., and Teusink, B. (2019). A systematic assessment of current genome-scale metabolic reconstruction tools. *Genome biology*, 20(1):1–20.
- Moretti, S., Martin, O., Van Du Tran, T., Bridge, A., Morgat, A., and Pagni, M. (2016). Metanetx/mnxref—reconciliation of metabolites and biochemical reactions to bring together genome-scale metabolic networks. *Nucleic acids research*, 44(D1):D523–D526.
- Mori, M., Hwa, T., Martin, O. C., De Martino, A., and Marinari, E. (2016). Constrained allocation flux balance analysis. *PLoS computational biology*, 12(6):e1004913.
- Nakashima, Y., Egami, Y., Kimura, M., Wakimoto, T., and Abe, I. (2016). Metagenomic analysis of the sponge discodermia reveals the production of the cyanobacterial natural product kasumigamide by ‘*entothionella*’. *PLoS One*, 11(10):e0164468.
- Nielsen, J. (2001). Metabolic engineering. *Applied microbiology and biotechnology*, 55(3):263–283.
- Nieselt, K., Battke, F., Herbig, A., Bruheim, P., Wentzel, A., Jakobsen, Ø. M., Sletta, H., Alam, M. T., Merlo, M. E., Moore, J., et al. (2010). The dynamic architecture of the metabolic switch in *streptomyces coelicolor*. *BMC genomics*, 11(1):1–9.

- Oberhardt, M. A., Zarecki, R., Gronow, S., Lang, E., Klenk, H.-P., Gophna, U., and Ruppin, E. (2015). Harnessing the landscape of microbial culture media to predict new organism–media pairings. *Nature communications*, 6:8493.
- O’Brien, E. J., Lerman, J. A., Chang, R. L., Hyduke, D. R., and Palsson, B. Ø. (2013). Genome-scale models of metabolism and gene expression extend and refine growth phenotype prediction. *Molecular systems biology*, 9(1):693.
- Olivares-Hernández, R., Bordel, S., and Nielsen, J. (2011). Codon usage variability determines the correlation between proteome and transcriptome fold changes. *BMC systems biology*, 5(1):33.
- Orth, J. D., Thiele, I., and Palsson, B. Ø. (2010). What is flux balance analysis? *Nature biotechnology*, 28(3):245–248.
- Øyås, O., Borrell, S., Trauner, A., Zimmermann, M., Feldmann, J., Liphardt, T., Gagneux, S., Stelling, J., Sauer, U., and Zampieri, M. (2020). Model-based integration of genomics and metabolomics reveals snp functionality in mycobacterium tuberculosis. *Proceedings of the National Academy of Sciences*, 117(15):8494–8502.
- Papoutsakis, E. T. and Meyer, C. L. (1985). Equations and calculations of product yields and preferred pathways for butanediol and mixed-acid fermentations. *Biotechnology and bioengineering*, 27(1):50–66.
- Patil, K. R., Åkesson, M., and Nielsen, J. (2004). Use of genome-scale microbial models for metabolic engineering. *Current opinion in biotechnology*, 15(1):64–69.
- Patil, K. R., Rocha, I., Förster, J., and Nielsen, J. (2005). Evolutionary programming as a platform for in silico metabolic engineering. *BMC bioinformatics*, 6(1):308.
- Reed, J. L. and Palsson, B. Ø. (2004). Genome-scale in silico models of e. coli have multiple equivalent phenotypic states: assessment of correlated reaction subsets that comprise network states. *Genome research*, 14(9):1797–1805.
- Reimers, A.-M., Knoop, H., Bockmayr, A., and Steuer, R. (2017). Cellular trade-offs and optimal resource allocation during cyanobacterial diurnal growth. *Proceedings of the National Academy of Sciences*, 114(31):E6457–E6465.
- Richards, M. A., Cassen, V., Heavner, B. D., Ajami, N. E., Herrmann, A., Simeonidis, E., and Price, N. D. (2014). Mediadb: a database of microbial growth conditions in defined media. *PLoS One*, 9(8):e103548.

-
- Rigali, S., Anderssen, S., Naômé, A., and van Wezel, G. P. (2018). Cracking the regulatory code of biosynthetic gene clusters as a strategy for natural product discovery. *Biochemical pharmacology*, 153:24–34.
- Roth-Rosenberg, D., Aharonovich, D., Luzzatto-Knaan, T., Vogts, A., Zoccarato, L., Eigemann, F., Nago, N., Grossart, H.-P., Voss, M., and Sher, D. (2020). Prochlorococcus cells rely on microbial interactions rather than on chlorotic resting stages to survive long-term nutrient starvation. *Mbio*, 11(4).
- Rügen, M., Bockmayr, A., and Steuer, R. (2015). Elucidating temporal resource allocation and diurnal dynamics in phototrophic metabolism using conditional fba. *Scientific reports*, 5:15247.
- Rutledge, P. J. and Challis, G. L. (2015). Discovery of microbial natural products by activation of silent biosynthetic gene clusters. *Nature reviews microbiology*, 13(8):509–523.
- Ryu, Y.-G., Butler, M. J., Chater, K. F., and Lee, K. J. (2006). Engineering of primary carbohydrate metabolism for increased production of actinorhodin in streptomyces coelicolor. *Applied and environmental microbiology*, 72(11):7132–7139.
- Saa, P. A. and Nielsen, L. K. (2016). Il-achrb: a scalable algorithm for sampling the feasible solution space of metabolic networks. *Bioinformatics*, 32(15):2330–2337.
- Saa, P. A. and Nielsen, L. K. (2017). Formulation, construction and analysis of kinetic models of metabolism: A review of modelling frameworks. *Biotechnology advances*, 35(8):981–1003.
- Salvy, P., Fengos, G., Ataman, M., Pathier, T., Soh, K. C., and Hatzimanikatis, V. (2019). pytfa and mattfa: a python package and a matlab toolbox for thermodynamics-based flux analysis. *Bioinformatics*, 35(1):167–169.
- Salvy, P. and Hatzimanikatis, V. (2020). Emergence of diauxie as an optimal growth strategy under resource allocation constraints in cellular metabolism. *bioRxiv*.
- Salzberg, S. L. (2019). Next-generation genome annotation: we still struggle to get it right. *Genome Biology*, 20(92).
- Sánchez, B. J., Zhang, C., Nilsson, A., Lahtvee, P.-J., Kerkhoven, E. J., and Nielsen, J. (2017). Improving the phenotype predictions of a yeast genome-scale metabolic model by incorporating enzymatic constraints. *Molecular systems biology*, 13(8):935.

- Sarkar, D., Mueller, T. J., Liu, D., Pakrasi, H. B., and Maranas, C. D. (2019). A diurnal flux balance model of *synechocystis* sp. pcc 6803 metabolism. *PLoS computational biology*, 15(1):e1006692.
- Sastry, A. V., Gao, Y., Szubin, R., Hefner, Y., Xu, S., Kim, D., Choudhary, K. S., Yang, L., King, Z. A., and Palsson, B. O. (2019). The *escherichia coli* transcriptome mostly consists of independently regulated modules. *Nature communications*, 10(1):1–14.
- Schellenberger, J. and Palsson, B. Ø. (2009). Use of randomized sampling for analysis of metabolic networks. *Journal of biological chemistry*, 284(9):5457–5461.
- Schnell, S. (2014). Validity of the michaelis–menten equation–steady-state or reactant stationary assumption: that is the question. *The FEBS journal*, 281(2):464–472.
- Schomburg, I., Chang, A., and Schomburg, D. (2002). Brenda, enzyme data and metabolic information. *Nucleic acids research*, 30(1):47–49.
- Schuetz, R., Kuepfer, L., and Sauer, U. (2007). Systematic evaluation of objective functions for predicting intracellular fluxes in *escherichia coli*. *Molecular systems biology*, 3(1):119.
- Schuetz, R., Zamboni, N., Zampieri, M., Heinemann, M., and Sauer, U. (2012). Multidimensional optimality of microbial metabolism. *Science*, 336(6081):601–604.
- Schuster, S. and Hilgetag, C. (1994). On elementary flux modes in biochemical reaction systems at steady state. *Journal of Biological Systems*, 2(02):165–182.
- Seemann, T. (2014). Prokka: rapid prokaryotic genome annotation. *Bioinformatics*, 30(14):2068–2069.
- Segre, D., Vitkup, D., and Church, G. M. (2002). Analysis of optimality in natural and perturbed metabolic networks. *Proceedings of the National Academy of Sciences*, 99(23):15112–15117.
- Shamir, M., Bar-On, Y., Phillips, R., and Milo, R. (2016). Snapshot: timescales in cell biology. *Cell*, 164(6):1302–1302.
- Sher, D., Thompson, J. W., Kashtan, N., Croal, L., and Chisholm, S. W. (2011). Response of *prochlorococcus* ecotypes to co-culture with diverse marine bacteria. *The ISME journal*, 5(7):1125–1132.

-
- Shlomi, T., Benyamini, T., Gottlieb, E., Sharan, R., and Ruppin, E. (2011). Genome-scale metabolic modeling elucidates the role of proliferative adaptation in causing the warburg effect. *PLoS Comput Biol*, 7(3):e1002018.
- Shlomi, T., Cabili, M. N., Herrgård, M. J., Palsson, B. Ø., and Ruppin, E. (2008). Network-based prediction of human tissue-specific metabolism. *Nature biotechnology*, 26(9):1003–1010.
- Stanway, R. R., Bushell, E., Chiappino-Pepe, A., Roques, M., Sanderson, T., Franke-Fayard, B., Caldelari, R., Golomngi, M., Nyonda, M., Pandey, V., et al. (2019). Genome-scale identification of essential metabolic processes for targeting the plasmodium liver stage. *Cell*, 179(5):1112–1128.
- Stark, R., Grzelak, M., and Hadfield, J. (2019). Rna sequencing: the teenage years. *Nature Reviews Genetics*, 20(11):631–656.
- Subramanian, A., Tamayo, P., Mootha, V. K., Mukherjee, S., Ebert, B. L., Gillette, M. A., Paulovich, A., Pomeroy, S. L., Golub, T. R., Lander, E. S., et al. (2005). Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proceedings of the National Academy of Sciences*, 102(43):15545–15550.
- Széliová, D., Ruckerbauer, D. E., Galleguillos, S. N., Petersen, L. B., Natter, K., Hanscho, M., Troyer, C., Causon, T., Schoeny, H., Christensen, H. B., et al. (2020). What cho is made of: Variations in the biomass composition of chinese hamster ovary cell lines. *Metabolic engineering*, 61:288–300.
- Szul, M. J., Dearth, S. P., Campagna, S. R., and Zinser, E. R. (2019). Correction for szul et al., “carbon fate and flux in prochlorococcus under nitrogen limitation”. *Msystems*, 4(2).
- Tepper, N. and Shlomi, T. (2010). Predicting metabolic engineering knockout strategies for chemical production: accounting for competing pathways. *Bioinformatics*, 26(4):536–543.
- Thiele, I., Fleming, R. M., Que, R., Bordbar, A., Diep, D., and Palsson, B. O. (2012). Multiscale modeling of metabolism and macromolecular synthesis in e. coli and its application to the evolution of codon usage. *PloS one*, 7(9):e45635.
- Thiele, I. and Palsson, B. Ø. (2010). A protocol for generating a high-quality genome-scale metabolic reconstruction. *Nature protocols*, 5(1):93.
- Thiele, I., Sahoo, S., Heinken, A., Hertel, J., Heirendt, L., Aurich, M. K., and Fleming, R. M. (2020). Personalized whole-body models integrate metabolism, physiology, and the gut microbiome. *Molecular systems biology*, 16(5):e8982.

- Thomas, L., Hodgson, D. A., Wentzel, A., Nieselt, K., Ellingsen, T. E., Moore, J., Morrissey, E. R., Legaie, R., Wohlleben, W., Rodríguez-García, A., et al. (2012). Metabolic switches and adaptations deduced from the proteomes of streptomyces coelicolor wild type and phop mutant grown in batch culture. *Molecular & Cellular Proteomics*, 11(2).
- Tyson, J. J. and Othmer, H. G. (1978). The dynamics of feedback control circuits in biochemical pathways. *Progress in theoretical biology*, 5:1–62.
- Ullah, E., Yosafshahi, M., and Hassoun, S. (2019). Towards scaling elementary flux mode computation. *Briefings in Bioinformatics*.
- van Dijk, E. L., Jaszczyszyn, Y., Naquin, D., and Thermes, C. (2018). The third revolution in sequencing technology. *Trends in Genetics*, 34(9):666–681.
- van Leeuwen, J., Pons, C., Tan, G., Wang, J. Z., Hou, J., Weile, J., Gebbia, M., Liang, W., Shuteriqi, E., Li, Z., et al. (2020). Systematic analysis of bypass suppression of essential genes. *Molecular systems biology*, 16(9):e9828.
- Varma, A. and Palsson, B. O. (1994a). Metabolic flux balancing: basic concepts, scientific and practical use. *Bio/technology*, 12(10):994–998.
- Varma, A. and Palsson, B. O. (1994b). Stoichiometric flux balance models quantitatively predict growth and metabolic by-product secretion in wild-type escherichia coli w3110. *Applied and environmental microbiology*, 60(10):3724–3731.
- Vaulot, D., Marie, D., Olson, R. J., and Chisholm, S. W. (1995). Growth of prochlorococcus, a photosynthetic prokaryote, in the equatorial pacific ocean. *Science*, 268(5216):1480–1482.
- Voigt, A., Nowick, K., and Almaas, E. (2017). A composite network of conserved and tissue specific gene interactions reveals possible genetic interactions in glioma. *PLoS computational biology*, 13(9):e1005739.
- Waldbauer, J. R., Rodrigue, S., Coleman, M. L., and Chisholm, S. W. (2012). Transcriptome and proteome dynamics of a light-dark synchronized bacterial cell cycle. *PloS one*, 7(8):e43432.
- Weber, T. and Kim, H. U. (2016). The secondary metabolite bioinformatics portal: Computational tools to facilitate synthetic biology of secondary metabolite production. *Synthetic and Systems Biotechnology*, 1(2):69–79.

-
- Wentzel, A., Bruheim, P., Øverby, A., Jakobsen, Ø. M., Sletta, H., Omara, W. A., Hodgson, D. A., and Ellingsen, T. E. (2012). Optimized submerged batch fermentation strategy for systems scale studies of metabolic switching in *Streptomyces coelicolor* a3 (2). *BMC systems biology*, 6(1):59.
- Wiback, S. J., Famili, I., Greenberg, H. J., and Palsson, B. Ø. (2004). Monte carlo sampling can be used to determine the size and shape of the steady-state flux space. *Journal of theoretical biology*, 228(4):437–447.
- Woodward, F. (2007). Global primary production. *Current Biology*, 17(8):R269–R273.
- Yang, L., Mih, N., Anand, A., Park, J. H., Tan, J., Yurkovich, J. T., Monk, J. M., Lloyd, C. J., Sandberg, T. E., Seo, S. W., et al. (2019). Cellular responses to reactive oxygen species are predicted from molecular mechanisms. *Proceedings of the National Academy of Sciences*, 116(28):14368–14373.
- Yang, L., Yurkovich, J. T., King, Z. A., and Palsson, B. O. (2018). Modeling the multi-scale mechanisms of macromolecular resource allocation. *Current opinion in microbiology*, 45:8–15.
- Zavřel, T., Faizi, M., Loureiro, C., Poschmann, G., Stühler, K., Sinetova, M., Zorina, A., Steuer, R., and Červený, J. (2019). Quantitative insights into the cyanobacterial cell economy. *Elife*, 8:e42508.
- Zhuang, K., Izallalen, M., Mouser, P., Richter, H., Risso, C., Mahadevan, R., and Lovley, D. R. (2011a). Genome-scale dynamic modeling of the competition between *Rhodospirillum rubrum* and *Geobacter* in anoxic subsurface environments. *The ISME journal*, 5(2):305–316.
- Zhuang, K., Vemuri, G. N., and Mahadevan, R. (2011b). Economics of membrane occupancy and respiro-fermentation. *Molecular systems biology*, 7(1):500.
- Zinser, E. R., Lindell, D., Johnson, Z. I., Futschik, M. E., Steglich, C., Coleman, M. L., Wright, M. A., Rector, T., Steen, R., McNulty, N., et al. (2009). Choreography of the transcriptome, photophysiology, and cell cycle of a minimal photoautotroph, *Prochlorococcus*. *PLoS one*, 4(4):e5135.
- Zomorodi, A. R., Islam, M. M., and Maranas, C. D. (2014). d-optcom: dynamic multi-level and multi-objective metabolic modeling of microbial communities. *ACS synthetic biology*, 3(4):247–257.
- Zur, H., Ruppin, E., and Shlomi, T. (2010). imat: an integrative metabolic analysis tool. *Bioinformatics*, 26(24):3140–3142.

Paper 1

Addressing Uncertainty in Genome-Scale Metabolic Model Reconstruction and Analysis

David B Bernstein, Snorre Sulheim, Eivind Almaas and
Daniel Segrè.


Genome Biology 22, 64 (2021)

REVIEW

Open Access



Addressing uncertainty in genome-scale metabolic model reconstruction and analysis

David B. Bernstein^{1†}, Snorre Sulheim^{2,3,4†}, Eivind Almaas^{3,5} and Daniel Segre^{1,2,6*} 

* Correspondence: dsegre@bu.edu

[†]David B. Bernstein and Snorre Sulheim contributed equally to this work.

¹Department of Biomedical Engineering and Biological Design Center, Boston University, Boston, MA, USA

²Bioinformatics Program, Boston University, Boston, MA, USA
Full list of author information is available at the end of the article

Abstract

The reconstruction and analysis of genome-scale metabolic models constitutes a powerful systems biology approach, with applications ranging from basic understanding of genotype-phenotype mapping to solving biomedical and environmental problems. However, the biological insight obtained from these models is limited by multiple heterogeneous sources of uncertainty, which are often difficult to quantify. Here we review the major sources of uncertainty and survey existing approaches developed for representing and addressing them. A unified formal characterization of these uncertainties through probabilistic approaches and ensemble modeling will facilitate convergence towards consistent reconstruction pipelines, improved data integration algorithms, and more accurate assessment of predictive capacity.

Introduction

Genome-scale metabolic models (GEMs) aim to capture a systems-level representation of the entirety of metabolic functions of a cell. They represent complex cellular metabolic networks using a stoichiometric matrix, which enables sophisticated mathematical analysis of metabolism at the whole-cell level [1]. Not only do GEMs provide a framework for mapping species-specific knowledge and complex ‘omics data to metabolic networks, but coupled with constraint-based reconstruction and analysis (COBRA) methods, such as Flux Balance Analysis (FBA), they facilitate the translation of hypotheses into algorithms that can be used to generate testable predictions of metabolic phenotypes [2–4]. These methods are now used to study biological systems for many different applications, including in metabolic engineering, human metabolism and biomedicine, and microbial ecology [5–11].

Over 100 well-curated GEMs exist for a range of prokaryotes and eukaryotes, offering an organized and mathematically tractable representation of these organisms’ metabolic networks [12, 13]. A detailed protocol has been described for the reconstruction of well-curated GEMs for new organisms [14]. Additionally, the increased



© The Author(s). 2021 **Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article’s Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article’s Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>. The Creative Commons Public Domain Dedication waiver (<http://creativecommons.org/publicdomain/zero/1.0/>) applies to the data made available in this article, unless otherwise stated in a credit line to the data.

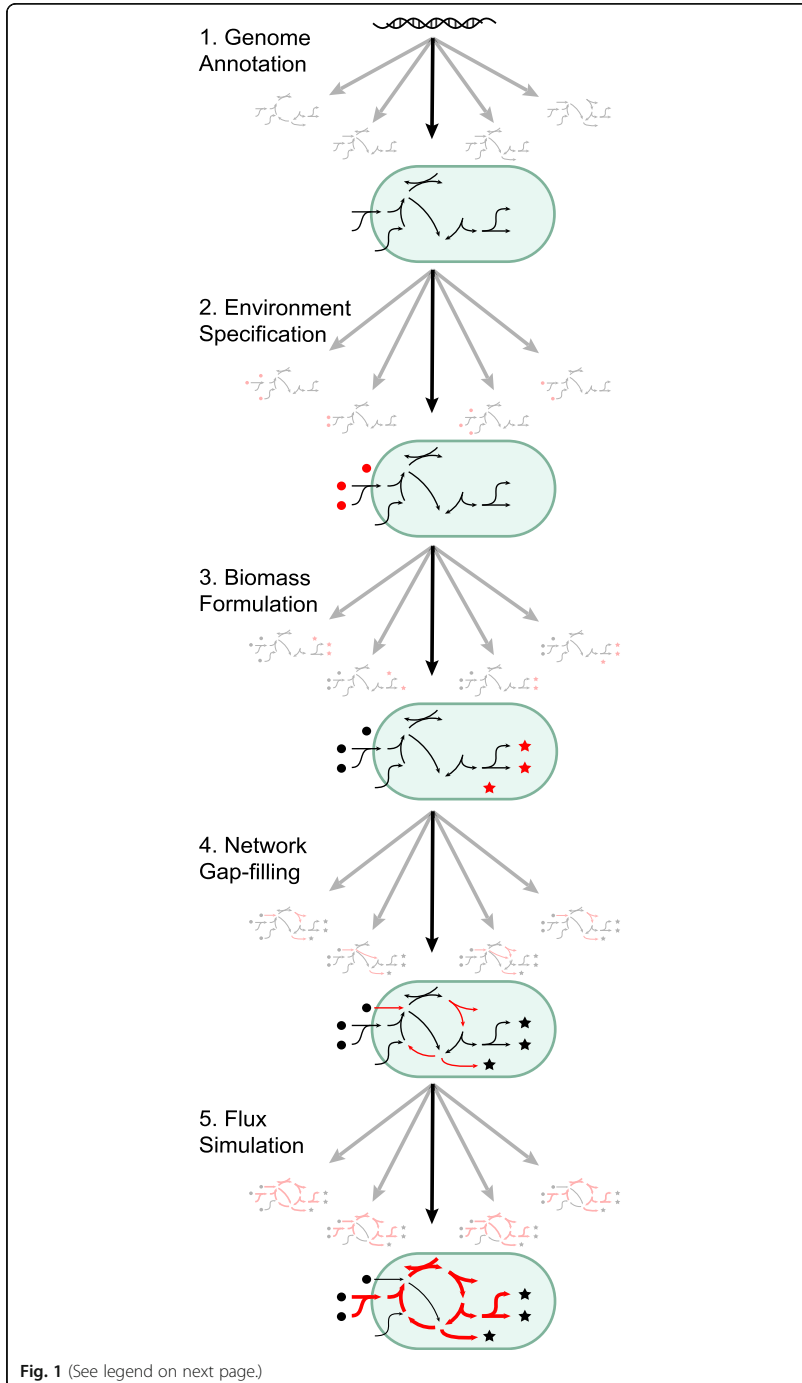
availability of whole-genome sequencing in combination with the development of pipelines for automatic model reconstruction has led to several frameworks that support rapid model reconstruction for a large number of non-model organisms [15–19]. For example, the US Department of Energy systems biology knowledgebase (KBase.us) currently enables the automatic generation of draft GEMs from over 80,000 sequenced genomes [20]. Thus, GEMs are rapidly becoming applicable for a wide range of biological applications.

Despite the numerous reconstructions and wide range of applications, GEMs have important limitations [21]. In this review, we focus on one major factor that currently limits the successful application of GEMs: the inherent uncertainty in GEM predictions that arises from degeneracy in both model structure (reconstruction) and simulation results (analysis). While GEM reconstructions typically only yield one specific metabolic network as the final outcome, this one network is indeed one of many possible networks that could have been constructed through different choices of algorithms and availability of information (Fig. 1). The process of GEM reconstruction is divided into (1) genome annotation, (2) environment specification, (3) biomass formulation, and (4) network gap-filling. Different choices in these first four steps can lead to reconstructed networks with different structures (reactions and constraints). On top of these choices, the final phenotypic prediction and biological interpretation is significantly affected by (5) the choice of flux simulation method. This review moves through these five different aspects of GEM reconstruction and analysis, outlining the key sources of uncertainty in each. In addition, we review various approaches that have been developed to deal with this uncertainty. We emphasize approaches that utilize probabilities or an ensemble of models to represent uncertainty. A table associated with each section outlines the different approaches that have been summarized and the sources of uncertainty that they address (Tables 1, 2, 3, 4 and 5).

Our ability to assess and communicate the sources of uncertainty associated with a model can have great impact on the relevance of predictions and on the degree to which these predictions can be constructively used for follow-up studies, as has been noted for the field of systems biology in general [22]. This review is not an introduction to genome-scale metabolic modeling or a survey of its applications, as these topics have been covered elsewhere [5, 11, 23]. Rather, we hope that this text will serve as a roadmap facilitating the development of methods that further formalize a unified characterization of uncertainty in GEM reconstruction and analysis.

Genome annotation

The first step towards a GEM reconstruction is the identification and functional annotation of the genes encoding metabolic enzymes present in the genome (Table 1). These annotations come from databases that employ homology-based methods for mapping genome sequences to metabolic reactions. The use of these annotation databases in GEM reconstruction pipelines in general is covered in several reviews [24–27]. It has been noted that the choice of a particular database significantly affects the structure of the reconstructed network [19]. This variability can be attributed to the limited accuracy of homology-based methods [28], misannotations present in large databases [29], the fact that many genes can only be annotated as hypothetical sequences of unknown function [30, 31], and the high fraction of “orphan” enzyme functions that



(See figure on previous page.)

Fig. 1 A general progression for genome-scale metabolic model reconstruction and analysis is represented by five major steps. The central black arrows demonstrate a standard approach, which yields a single output from each step. The gray arrows represent the uncertainty in this process, with the output of each step as an ensemble of possible results. The new additions to the model at each step are shown in red: circles represent metabolites, stars represent biomass components, arrows represent metabolic reactions, and bold arrows represent a specific flux distribution

cannot be mapped to a particular genome sequence [32]. Some, but not all, of this variability can be mitigated by combining multiple databases to increase the coverage of annotation when reconstructing a GEM [33, 34]. Furthermore, annotation for GEM reconstruction has an added layer of complexity beyond mapping genes to general ontologies or homologs. It is necessary to map genes to the metabolic reactions that they enable. These mappings, referred to as gene-protein-reaction association rules, use Boolean expressions to encode the nonlinear mapping between genes and reactions (manifested in multimeric enzymes, multifunctional enzymes and isoenzymes). The reconstruction and interpretation of these rules adds additional uncertainty to the annotation process. Even if a rule faithfully represents the functional possibilities encoded in a set of genes, the cellular “interpretation” of the rule may be highly nuanced and complex. For example, isoenzymes may not always compensate for each other’s deletion due to different regulatory couplings [35], and alternative usage of the Boolean relationship may best capture the cost of a gene deletion and its degree of evolutionary conservation [36]. An innovative approach for representing gene-protein-reaction association rules is to encode them into the stoichiometric matrix of the GEM [37]. This encoding makes it possible to extend flux sampling approaches to gene sampling, facilitating the quantification of uncertainty. These sampling approaches are discussed further in the flux simulation section.

A few reconstruction pipelines try to circumvent the problem of incorrect or missing functional annotation by using previously curated GEMs as annotation templates. Using several different reconstruction pipelines—RAVEN [38, 39], AuReMe/Pantograph [40, 41], or MetaDraft [42]—the user can map annotations from one organism directly to a curated model of a closely related organism by employing homology searches between the two. In this way, well-curated metabolic reaction annotations from an established GEM are propagated to new GEM reconstructions. Another reconstruction pipeline, CarveMe, uses a curated network of all possible reactions, based on the BiGG database [13], as the reference and “carves out” a subset of reactions to create organism-specific models [43]. While these methods may provide more complete reconstructions that require less gap-filling, they do not solve the fundamental issue of the uncertainty in the mapping of homologs or provide an estimate of the uncertainty associated with the presence of each reaction in the network.

Another approach is to directly incorporate uncertainty in functional annotation by assigning several likely annotations to each gene rather than picking the single most likely. In one likelihood-based approach, metabolic reactions are annotated probabilistically by taking into account the overall homology score, BLAST e-value, and keeping track of suboptimal annotations [44]. In this approach, metabolic reactions are assigned a probability of being present in a GEM based on both the strength and the uniqueness of the annotation. This approach has been developed into the ProbAnnoPy and

ProbAnnoWeb pipelines that provide probabilistic annotations in the ModelSEED framework [45]. Beyond using only homology from BLAST to inform annotation probabilities, the CoReCo algorithm has additionally included homology scores based on global trace graphs, which have been proposed as an improved approach for identifying distant homologs [46]. The CoReCo algorithm also utilizes phylogenetic information to improve the probabilistic annotation of GEMs for multiple organisms simultaneously. Additional context information has also been incorporated into a probabilistic metabolic reaction annotation approach in the GLOBUS algorithm [47]. Context-based information includes gene correlations from transcriptomics, co-localization of genes on the chromosome and phylogenetic profiles, all of which are complementary to gene-sequence homology for inferring functional protein annotations. The probabilistic metabolic reaction annotations generated with these methods serve as a good starting point for subsequent reconstruction steps. For example, the likelihood-based approach mentioned here is used to implement a probabilistic gap-filling algorithm, further discussed in the gap-filling section [44].

Other concepts that have been used to generally improve gene functional annotation could be further incorporated into GEM annotation pipelines. For example, functional annotation of enzymes could be improved by the incorporation of enzyme active/catalytic site information from databases such as M-CSA [48]. Additionally, the annotation of specific classes of proteins, such as biosynthetic gene clusters [49, 50], transporters [51, 52], and amino acid biosynthetic pathways [53], can be improved by using approaches tailored to identify features that are specific to those protein classes. In particular, transport reactions are difficult to properly annotate and can add significant uncertainty to GEMs [14]. For example, the substrate specificity of automatically annotated transport reactions can often be improved with experimental data [54]. Furthermore, incorrect transport reactions can cause ATP generating cycles that lead to inaccuracies in GEM predictions [55]. Beyond traditional annotation approaches, machine learning has also been used to improve enzyme annotation by predicting EC numbers directly from gene sequences, potentially picking up on subtle features that would otherwise be missed by homology-matching-based approaches [56]. The localization of reactions to specific compartments is an added layer of annotation that is important for accurate GEM reconstruction, especially of eukaryotes [57, 58]. Also in this case, machine learning approaches can be used to predict the specific subcellular localization of proteins [59, 60]. New high-throughput genomics experimental methods can also be used to simultaneously assess the function of many genes in a large number of environments [54, 61]. Incorporating novel ideas from these methods into GEM reconstructions may reduce the overall uncertainty of functional annotation.

Environment specification

To use a GEM for the prediction of expected phenotypes, or for the simulation of dynamic processes, one must define the chemical composition of the environment (Table 2). Establishing the list of environmentally available molecules is straightforward in simple laboratory experiments, in which defined media with known chemical composition are used. In this context, databases such as Media DB [62] or KOMODO [63] have cataloged a large number of defined media, greatly facilitating metabolic modeling. Many laboratory experiments, however, are performed in undefined media containing

Table 1 Summary of approaches that address sources of uncertainty in genome annotation. Highlighted in bold are key approaches related to probabilistic or ensemble-based methods

Approach	Sources of uncertainty	References
Comparison of pipelines	Variability across databases	[19]
Combining databases	Variability across databases	[33, 34]
Template GEMs	Incomplete annotations in non-model organisms	[38–43]
Probabilistic annotation	Annotation errors	[44, 45]
Probabilistic annotation + context Information	Annotation errors	[46, 47]
Specific databases and high-throughput genomics	Annotation errors	[48–54, 56, 59–61]

ingredients such as “yeast extract” that cannot be easily listed and quantified. In nature, microbes often exist in highly complex environments where the chemical inputs to the system are undefined, vary with time, and are altered by other microbes in the environment. Furthermore, it is not sufficient to know the list of compounds present in the cultivation medium, but one must also know at what rates the compounds can be consumed by the organism to properly set the bounds on the uptake reactions of the metabolic model. In principle, the composition of the environment can be determined through experimental techniques such as exo-metabolomics, where measurements of metabolites in the extracellular environment are used to infer cellular uptake and secretion rates [64–68]. This approach can provide valuable information for reducing the uncertainty in the environment specification. However, this data comes with its own uncertainty that should be carefully addressed [69]. All of these factors lead to a wide range of uncertainty arising in environment specification for metabolic network analysis [70].

GEMs provide an opportunity to address the uncertainty associated with complex environments. GEM analysis algorithms, such as FBA, are computationally efficient and can thus be run across a large ensemble of environments to quantify the sensitivity of simulated fluxes to nutrient composition. Several studies have quantified this sensitivity by identifying aspects of GEM predictions that are either strongly affected by or robust to variation in the environmental composition [71–77]. Describing this sensitivity, or robustness, provides a clearer picture of how uncertainty in the environment specification may, or may not, propagate to specific GEM predictions. Early on, phenotype phase plane analysis was developed to show the impact on optimal growth rate of varying the fluxes of two limiting resources [71, 72]. Moving beyond pairs of resources, large ensembles of nutrients can be randomly sampled to assess the variability of all intracellular fluxes. For example, Almaas et al. showed, using a well-curated *Escherichia coli* GEM, that the overall distribution of metabolic fluxes is robust to the environmental composition; however, specific fluxes vary, with most discrete variations occurring in a connected “high-flux backbone” of reactions [73]. Subsequent work highlighted the evolutionary importance of an active core of reactions that carry flux in all environments [74]. Reed and Palsson further demonstrated that reactions with correlated fluxes across environments are indicative of transcriptional regulatory structure [75]. These studies point to the non-trivial nature of the sensitivity of GEM predictions to

environment specification. Beyond the context of individual organisms, GEM analysis has been used to demonstrate that varying the environment can alter the nature of metabolic interactions between microbial organisms [78] and that certain environmental variables, such as the presence of oxygen, can have a significant impact on the interaction types that arise [79]. Variable environments can impact cellular metabolism from individual reaction fluxes up to the level of microbial interactions. Thus, in applications where the environment is uncertain, ensemble or probabilistic approaches are needed to fully capture potential phenotypes.

A more recent approach, inspired by the statistical physics concept of network percolation, utilizes random sampling of nutrient compositions to quantify which metabolites can be consistently produced by a given metabolic network across many environments [80]. This approach introduced a probabilistic framework for representing the input metabolites of a metabolic network, which could further facilitate random sampling of environmental ensembles in future methods. While the current implementation of this framework samples all environmental metabolites with equal probability, one could envisage future approaches which represent environmental uncertainties more accurately by using biased distributions that incorporate any available knowledge. This approach would fill the existing gap between assuming a single known environment and randomly sampling environments uniformly. Additionally, environment sampling could be used to vary the flux (in FBA) or concentrations (in dynamic FBA) of different environmental components, in addition to their presence and absence, to assess the impact of these quantities on metabolic network properties.

The specification of the environment for GEM analysis could be further improved using “reverse ecology” methods that aim to infer the native environment from the metabolic network structure either through constraint-based optimization [81–83] or by defining “seed” metabolites that are needed as inputs for a metabolic network and are therefore more likely to be found in that organism’s natural environment [84, 85]. Since these methods utilize the metabolic network structure to inform the environment specification, they should be applied carefully as uncertainty in the network may propagate into environment specification.

Biomass formulation

The cell biomass used in GEMs is an inventory list of all compounds essential for growth of a given organism, weighted to represent the amount of each component present in 1 g of dry-weight biomass. The reaction that transforms all biomass components into a unit of biomass is used to represent growth in GEMs and is necessary to

Table 2 Summary of approaches that address sources of uncertainty in environment specification. Highlighted in bold are key approaches related to probabilistic or ensemble-based methods

Approach	Sources of uncertainty	References
Media databases	Inconsistent media definition	[62, 63]
Experimental determination	Undefined environment composition	[64–68]
Phenotype phase plane	Variable environment composition	[71, 72]
Ensemble sampling	Variable environment composition	[73–79]
Probabilistic sampling	Variable environment composition	[80]
Reverse ecology	Undefined environment composition	[81–85]

perform popular analyses such as FBA. Since several aspects of the biomass reaction and its use have been reviewed before [86], we will focus on the uncertainty associated with its formulation (Table 3).

The main source of uncertainty in the formulation of biomass composition is the lack of direct experimental measurements for most organisms. In the absence of specific data, the biomass composition from a model organism (e.g., *E. coli* for Gram-negative or *Bacillus subtilis* for Gram-positive bacteria) is often used as template, despite the significant uncharacterized variation in biomass composition likely to exist across different organisms. This trend has been verified by hierarchical clustering of biomass compositions from 71 curated GEMs: rather than taxonomic relations, the clusters were defined by the template biomass functions used in the model reconstruction [87]. Similarly, in a survey of plants, the biomass was only experimentally determined in 5 of 21 GEMs [88]. Furthermore, even within the same organism, the biomass composition can change in response to changes in growth rate, nutrient availability, temperature, and osmotic stress [89–95].

A number of studies have addressed the sensitivity of model predictions to changes in biomass formulation. Because these studies differ both in how the biomass function is changed and which model predictions are evaluated, they reach different conclusions. Initially, Pramanik and Keasling used correlations between growth rate and macromolecular abundances to estimate growth-rate-specific biomass compositions in *E. coli* [96, 97]. When the high growth-rate biomass composition was used to simulate fluxes in a low growth-rate environment, or vice versa, the total deviation from measured fluxes increased drastically compared to simulations with correct biomass specification [96]. Secondly, they showed that the predicted fluxes were sensitive to quantitative changes in the fatty acid composition of the biomass [97]. More recent analyses of the effect of changing the biomass composition in *Saccharomyces cerevisiae* have shown large influence on gene knock-out growth predictions [98], variable effect on substrate uptake rates [99], and an effect on the flux distribution dependent on the identity of the limiting nutrient [100]. In contrast, little effect was found on the predicted growth yield in *Pseudomonas putida* [101]. To address the dependence of the biomass formulation on the environment, within an individual organism, Schulz et al. propose two concepts for the incorporation of, or interpolation between, multiple biomass functions corresponding to different growth environments [102]. The first concept allows the GEM to choose an optimal linear combination of existing biomass functions while the second concept uses a hyperplane interpolation to predict the correct biomass function for the selected growth environment. The authors use hypothetical biomass functions to show that the choice of method has a clear impact on model predictions, but further evaluation calls for experimental follow-up. Swapping the biomass between different organisms can provide insight into the sensitivity of GEMs to strain specific biomass formulations, which is an important consideration given the widespread use of template biomass formulations. Leveraging three independent reconstructions of *Arabidopsis thaliana* with substantially different biomass reactions, it was found that the fluxes in central carbon metabolism were robust to replacement of the biomass reaction from one of the other models [88]. In contrast, swapping biomass reactions between five different bacterial species resulted in up to 30% change in predicted essential reactions [87].

Although the effect of uncertainty and error in the biomass coefficients depends on a large number of variables and how the effect is measured, it is clear that GEMs would benefit from increased precision in the estimation of biomass coefficients, which would ideally be organism and condition specific. The need for accurate estimates of the biomass composition has recently been addressed by experimental protocols [103–105] and the software BOFdat [106]. BOFdat provides a pipeline for computation of biomass coefficients and reports that the macromolecular composition is the most important factor in determining stoichiometric coefficients and should therefore be prioritized above ‘omics datasets. One elegant feature of BOFdat is a genetic algorithm which samples ensembles of biomass formulations to identify carbohydrate and small-molecule compositions such that model simulations optimally correspond with knock-out phenotype data. Looking forward, approaches such as BOFdat could be used to represent uncertainty in the biomass composition by sampling from an ensemble of possible biomass equations. Likewise, uncertainty in the stoichiometry of each biomass component could be incorporated by probabilistically sampling each coefficient from an appropriate distribution. Experimental data could be incorporated into this process to guide and constrain the distributions that are sampled through a Bayesian approach.

Network gap-filling

Gap-filling is an important step in GEM reconstruction that transforms a draft network into one that can produce biomass in the specified environment (Table 4). The idea of gap-filling—that missing knowledge in metabolism may require algorithms to identify reactions absent in the representation of a specific pathway, but likely present in the organism—has been around since the early days of metabolic network modeling [107]. Gap-filling algorithms in general have been reviewed previously [108], but in brief, they utilize a universal database of possible reactions to augment an existing metabolic network with the goal of enabling feasible growth states, e.g., by connecting dead-end metabolites. Here we focus on the uncertainty associated with this process. Gap-filling is inherently uncertain because the reactions added are generally not supported by genomic evidence. Moreover, multiple solutions can often be found to satisfy the same gap-filling problem. Due to this uncertainty, basic gap-filling algorithms are known to be somewhat inaccurate [109], prompting recent benchmarking on randomly degraded metabolic networks to highlight the variability in gap-filling performance [110]. Furthermore, many GEMs contain significant inconsistencies even after the application of gap-filling approaches, and their identification is important for ensuring model fidelity [111].

Table 3 Summary of approaches that address sources of uncertainty in biomass formulation. Highlighted in bold are key approaches related to probabilistic or ensemble-based methods

Approach	Sources of uncertainty	References
Alternative biomass formulations	Variability in biomass within organisms	[96–101]
Environment-dependent biomass formulation	Variability in biomass within organisms	[102]
Cross-organism biomass comparison	Biomass differences across organisms	[87, 88]
Experimental determination	Undefined biomass composition	[103–105]
Ensemble sampling	Undefined biomass composition	[106]

The uncertainty in gap-filling solutions has prompted the development of various probabilistic approaches to integrate data and prioritize solutions. An early innovation in probabilistic gap-filling algorithms was the development of a method to evaluate the addition of reactions to fill gaps based on a Bayesian network including sequence homology, operon, and pathway-based information [112]. A similar approach is to use probabilistic weights during the gap-filling process, such that more probable reactions incur a smaller penalty when added to the metabolic network. The CROP algorithm is an example of gap-filling based on growth phenotype data that implements weights based on various sources of evidence, including manually curated experimental evidence, pathways known to be associated with an organism, thermodynamics, and probabilistic estimates of enzyme function [113]. Another probabilistic approach has been developed to translate sequence homology into the likelihood that a metabolic reaction is present in a given metabolic network (discussed in the “Genome annotation” section); these likelihoods can then be used as probabilistic weights during the gap-filling procedure [44, 45].

Beyond probabilistic gap-filling methods, ensemble approaches have been developed to represent the uncertainty in gap-filling solutions as an ensemble of possible gap-filled GEMs. An early approach in this area prunes a universal metabolic network to identify locally minimal gap-filling solutions that align with experimental data [114]. In this approach, an ensemble of metabolic networks is generated by randomly assigning the order in which reactions are pruned from an original universal metabolic network. A similar pruning-based ensemble method, MIRAGE, additionally includes gene expression and phylogeny when weighting the order in which to remove reactions [115]. The idea of ensemble gap-filling was more fully developed by an approach that utilizes growth phenotype data in a randomized order to generate an ensemble of gap-filling solutions [116]. By randomly changing the sequence in which growth phenotype data was presented to the gap-filling algorithm, Biggs and Papin generated an ensemble of metabolic networks that equally agree with the given data. This study further demonstrated that utilizing the ensemble gap-filling result can be more accurate than using the individual results, or a global simultaneously gap-filled result. An additional ensemble gap-filling approach is implemented in the CarveMe method. CarveMe generates ensembles of gap-filled models by assigning random weights to reactions without genomic evidence [43].

Finally, automated gap-filling methods are fundamentally limited by the underlying database(s) of metabolic reactions that they utilize [117, 118]. Thus, uncertainty in this database set can have a large impact on gap-filling performance. This is a major limitation when considering the complexity of the true metabolic universe and the fact that we likely do not know the proper annotations for all metabolic reactions. In light of this limitation, a number of methods have been developed to predict possible metabolic reactions based on general reaction rules. Many of these approaches have been reviewed previously in the context of predicting biosynthetic pathways for target compounds [25, 119, 120]. One of the earlier approaches, the BNICE framework, expands the metabolic universe by learning generic reaction rules from the KEGG reactome [121]. This framework was subsequently used to develop MINE and ATLAS, databases of theoretically possible compounds and enzymatic reactions, respectively [122–124]. BNICE also suggests three-level EC-numbers for hypothetical reactions, which can guide discovery of proteins associated with de novo reactions. The theoretical number of reactions in the

expanded ATLAS is more than 10-fold higher than the number of reactions in KEGG, indicating that a large number of unexpected chemical transformations may be involved in metabolism. As we grapple with uncertainty in metabolic network reconstruction, de novo methods such as these can help us address unknown unknowns and provide exciting unanticipated insights. Moving forward, a combination of probabilistic and ensemble methods for data integration and de novo reaction prediction will enable the generation of gap-filled metabolic networks that represent uncertainty and can be better used to guide model refinement.

Flux simulation

One of the most common and powerful uses of GEMs is the prediction of metabolic phenotypes at steady state through the computation of expected fluxes through each reaction. Because the rank of the stoichiometric matrix is almost always less than the number of reactions, the linear system of equations associated with steady state is, in general, underdetermined. Thus, there are an infinite number of solutions within the multidimensional solution space (a space where each dimension corresponds to the flux of a metabolic reaction) [125]. Any point within the solution space is a feasible solution representing a metabolic phenotype. While there often is an emphasis on identifying the *correct* solution in this solution space (i.e., an individual point closest to the outcome of experimental measurements), choices and uncertainty in some of the above aspects of the computation necessarily lead to uncertainty in the prediction of the fluxes themselves. In this section, we will review prior work addressing this uncertainty, with an emphasis on methods geared towards embracing and reporting it (Table 5).

The flagship method for simulating metabolic fluxes in GEMs, FBA, uses linear programming to identify a point (or a subspace) in the solution space that optimizes a predefined cellular objective [23, 126–129]. Quite often, this objective is chosen to be the maximization of biomass production. A fundamental question that has surrounded the FBA approach since its early days is whether and under what conditions the assumption that biological systems operate close to a predictable optimum is valid, and if so, which objective function best represents the metabolic goals of a cell. Several studies have explored this uncertainty associated with the choice of the objective function. Schuetz et al. show that intracellular fluxes can be accurately predicted using FBA and an appropriate cellular objective [130]. However, none of the 11 selected objectives could provide the best predictability across different conditions when comparing predicted fluxes with ^{13}C flux experiments in *E. coli*. It was early on demonstrated that FBA with maximization of growth rate could predict the phenotype of *E. coli* wild-type strains, supporting the assumption that unicellular organisms have evolved towards

Table 4 Summary of approaches that address sources of uncertainty in network gap-filling. While all gap-filling approaches address uncertainty arising from missing annotations, here we point out approaches that address uncertainty in the gap-filling solutions. Highlighted in bold are key approaches related to probabilistic or ensemble-based methods

Approach	Sources of uncertainty	References
Evaluating gap-filling accuracy	Degenerate solutions	[109, 110]
Probabilistic gap-filling	Degenerate solutions	[44, 45, 112, 113]
Ensemble gap-filling	Degenerate solutions	[43, 114–116]
De novo reaction prediction	Reaction database incompleteness	[121–124]

maximal growth [131]. Indeed, by minimizing the deviation from measured fluxes in yeast, maximization of growth rate was identified as the most likely objective in glucose-limited conditions [132]. Taking an inverse FBA approach, Zhao et al. predicted the objective function for *E. coli* strains evolved through 50,000 generations [133]. Although they identified an infinite number of objective functions that could describe the measured flux ratios, maximization of biomass alone was not one of these objectives [134]. A different study of these *E. coli* strains also provided nuance to our understanding of evolutionary pressures by confirming that *E. coli* evolves towards maximization of growth rate primarily by increasing substrate usage, but only if the ancestral strain is initially far from the optimum [135].

In a number of instances, the phenotypes of knock-out mutants are actually more accurately predicted when taking into account suboptimal solutions (near but not exactly on the FBA predicted optimum). For example, the increased accuracy of the MOMA and related methods stems from the assumption that a knock-out strain is still steered towards the wild-type optimum by the cellular regulatory network and may not necessarily approach the knock-out optimum [136]. The PSEUDO method can further improve the accuracy of knock-out flux predictions by assuming that the knock-out flux is closest to a degenerate space of suboptimal solutions near the wild-type optimum, representing regulatory variability around the wild-type solution [137]. The optimality of solutions has been further investigated in a study leveraging ^{13}C -measurements of 9 different bacteria, which found that metabolism operates close to a Pareto surface that balances the trade-off between maximization of growth and ability to adapt to changing conditions [138]. In summary, these results suggest that suboptimality may provide increased robustness to stochastic variation and perturbation, a property with known importance in biological systems [139, 140].

To avoid biased assumptions of the metabolic goal of a microorganism, one can characterize the complete solution space to describe all possible phenotypes satisfying the steady-state and flux constraints. It is important to note that, even at the optimum predicted by FBA, the solution is rarely unique. The predicted flux vector must therefore be analyzed with caution. Flux variability analysis (FVA) can be used to estimate the range of possible fluxes at the optimum [141], but since the range of each reaction is estimated independently, the method provides no information on the correlations between fluxes. More sophisticated methods include enumeration of alternative optima [142–145], or a full description of the solution space through flux coupling [146], extreme pathway analysis [147], elementary flux modes (EFMs) [148], and elementary flux vectors (EFVs) [149]. EFMs decompose the steady-state solution space into characteristic support minimal vectors, while EFVs have the added benefit of incorporating flux bounds to further constrain the space to a polyhedron. Although these methods provide an unbiased framework for identifying metabolic pathways, a representation of the entire solution space is generally intractable for genome-scale models because of the non-polynomial scaling with the number of reactions [150].

Random sampling provides a scalable approach to describe possible phenotypes in the solution space. Monte-Carlo-based algorithms [151–153] have proven useful for a large number of applications [154], from a general description of the distribution of metabolic fluxes [73, 155, 156] to transcriptional regulation of key enzymes [157] or comparison of bacterial strains [158]. However, verification of convergence is a key

quality control of random sampling results currently lacking in analysis of GEMs [159]. The computational time required to reach convergence is a practical issue for large models, but recent work shows that the sampling results can be estimated at a reduced cost by using analytical methods and Bayesian inference [160]. Random sampling of the flux space can also be probabilistically biased to better represent uncertainty. A recent concept estimates the probability distribution of flux states that maximizes entropy with an average growth rate equal to the experimental value [161, 162]. As stated in the principle of maximum entropy, this probability distribution is the best representation of available knowledge [163, 164]. Another recently developed approach, Bayesian FBA, can be used to sample metabolic fluxes from a truncated multivariate normal distribution with prior distribution centered around zero [165]. In Bayesian FBA, prior knowledge such as measured growth and uptake rates, or ^{13}C - flux data, can be elegantly incorporated in calculations of posterior flux distributions in a generic Bayesian framework that provides insight into the uncertainty associated with individual fluxes and flux couplings.

The uncertainty in model predictions can be reduced by introduction of additional constraints which reduce the size of the solution space [3, 125]. The most common constraints are those associated with limits on nutrient uptake (as defined by the environment composition), thermodynamic irreversibility, and the presence of specific reactions, such as the growth and non-growth associated maintenance [166, 167]. However, these constraints have their own associated uncertainties. Uncertainty in growth and non-growth associated maintenance derives both from the experimental growth data used to estimate these values [14], and variability in the maintenance cost of cellular processes in different environments and organisms [168]. The impact of this uncertainty on GEM predictions has only been briefly touched upon [169, 170]. Taking into account thermodynamic constraints on metabolic reaction fluxes is a powerful approach to improve model predictions, both by identifying subnetworks violating the second law of thermodynamics and to infer the direction of metabolic reactions from the calculated change in Gibbs free energy [55, 171–174]. However, the calculation of Gibbs free energy for the large number of reactions present in GEMs requires approximate approaches, such as the group contribution method [175, 176].

Another branch of methods uses either transcriptome [177] or proteome [178, 179] data to constrain reaction fluxes according to the abundance of proteins catalyzing the respective metabolic reactions. While transcriptomics data have the benefit of increased coverage of genes compared to proteomics (e.g., covers 60% of the enzymes in the yeast-GEM) [178], the transcript levels do not necessarily correlate with enzyme abundance [180, 181]. This may explain why Parsimonious enzyme usage FBA (pFBA), which minimizes the total sum of the absolute values of fluxes [182], in general outperformed seven different transcriptome-based methods in predicting intracellular fluxes for both *S. cerevisiae* and *E. coli* across three different conditions [177]. An additional advantage of pFBA is that it does not require additional parameters, unlike the aforementioned transcriptomics/proteomics approaches, which may require a large number of parameters to properly integrate the data. Similar to pFBA, several other methods use global constraints to improve model predictions. Of particular interest are Constrained Allocation Flux Balance Analysis (CAFBA) [183] which takes the growth-dependent ribosome allocation into account, the global constraint of dissipation of

Gibbs free energy [184], and the extension of pFBA to include reaction likelihoods [185]. In any of these methods, particularly those that use additional data and parameters, it is important to remember that additional data used to further constrain the flux space comes with its own associated uncertainty, which must be taken into account when integrating it into GEMs.

The steady-state assumption forms the basis of constraint-based analysis by requiring mass-balance of all intracellular metabolites and defines the solution space discussed throughout this section. This assumption is justified because transient changes in metabolite concentrations occur rapidly compared to environmental and regulatory perturbations, leading to rapid convergence to a quasi-steady-state where metabolite concentrations are constant [186, 187]. However, when considering the uncertainty in stoichiometric coefficients, particularly in the biomass function, the steady-state assumption is effectively relaxed [165, 188, 189]. The RAMP approach demonstrates that relaxing the steady-state assumption can lead to more accurate predictions of intracellular fluxes [189]. The RAMP solution converges to the FBA solution when the uncertainty in stoichiometric coefficients approaches zero, demonstrating that this is a more general approach. While only uncertainty in the coefficients of the biomass reaction is explicitly tested in this work, RAMP's general framework is not limited to this case and can include uncertainty in reaction bounds or uncertainty in coefficients associated with protein allocation or thermodynamics.

Discussion

In this review, we highlighted methods that use probabilistic approaches and ensemble modeling to represent the uncertainty associated with constraint-based reconstruction and analysis of GEMs. Formalizing the representation of uncertainty in GEMs would improve confidence in modeling results. Although we concede that this is a difficult task, we hope that this review will serve as a roadmap for how this issue can be further addressed. We maintain that ensemble approaches (which are in essence discrete representations of probability distributions) provide a strong framework that naturally captures the uncertainty arising from the many possible outcomes in each step of the reconstruction and flux analysis process (Fig. 1). A practical step moving forward is the development of a unified metabolic network reconstruction and analysis framework that provides a probabilistic ensemble of results. Such a framework would require further development of methods for the representation and analysis of GEM ensembles,

Table 5 Summary of approaches that address sources of uncertainty in flux simulation. Highlighted in bold are key approaches related to probabilistic or ensemble-based methods

Approach	Sources of uncertainty	References
Alternative objective functions	Undefined cellular objective	[130–132, 134, 135]
Suboptimal solutions	Undefined cellular objective	[136–138]
Characterization of optimal solutions	Degenerate optimal solutions	[141–145]
Characterization of steady-state solution space	Degenerate solution space	[146–149]
Random sampling	Degenerate solution space	[151–160]
Random sampling with probabilistic biases	Degenerate solution space	[161, 162, 165]
Added constraints	Degenerate solution space	[55, 168–174, 177–179, 182–185]
Relaxed steady-state assumption	Steady-state assumption	[188, 189]

such as the MEDUSA package [190], and continued development and integration of approaches that represent uncertainty encountered in each stage of the GEM reconstruction and analysis process. In future development of ensemble models of GEMs, one should keep in mind that this approach is not a panacea [191]. It will be important to accurately account for uncertainty in each step to avoid potential pitfalls, such as an increase in false positive predictions given the sparse nature of the stoichiometric matrix. For example, when incorporating de novo predicted reactions into network gap-filling algorithms, the probabilistic weighting of these reactions would need to be carefully tuned. Additionally, it will be important to further explore correlations between the results of the different steps in the reconstruction and analysis process to fully understand uncertainty in this framework. For example, probabilistic genome annotation and ensemble gap-filling can work synergistically to identify candidate genes for orphan metabolic reactions. Conversely, uncertainty in metabolic network structure could be propagated through methods that use the network structure to infer the biomass formulation (such as BOFdat) or environment specification (such as reverse ecology). It is also important to focus on understanding the sensitivity of modeling results to uncertainty in specific parameters or steps in the pipeline. Generating an ensemble of results can provide insight into which results are robust to uncertainty in different parameters or model choices. Furthermore, clustering and classifying ensembles of results with machine learning algorithms can provide insight into which areas of genome-scale modeling are particularly sensitive and should be targeted for uncertainty reduction [192]. Ultimately, capturing all of the uncertainty in GEM reconstruction and analysis in a single pipeline will be a difficult task, and an emphasis should be placed on transparency and reproducibility such that all of the assumptions employed by a particular approach can be easily accounted for [193]. The standardization of model quality control provided by MEMOTE is an important contribution in this direction [194]. A similar community-effort towards standardized assessment and reporting of GEM uncertainties, as has been recently suggested by Carey et al., would be similarly highly beneficial [195].

Multomics data integration is an increasingly important application of GEMS as biological studies are now collecting and analyzing multiple sources of high-throughput data. GEMs can facilitate the integration of this data in a knowledge-based format that provides mechanistic insight [20, 196]. Approaches and challenges in integrating 'omics data into GEMs have been reviewed previously, with a particular focus on the difficulty of precise data integration due to GEMs' lack of kinetic information [197]. It is important to consider how best to represent 'omics data such that they can be integrated into GEMs. In line with the main message of our review, Ramon et al. suggest that a Bayesian perspective can aid the integration of 'omics data by taking into account the uncertainty in the metabolic network and experimental observations [197]. In this context, 'omics data can be used to constrain both the prior and posterior distributions from which ensembles of GEMs are sampled. Furthermore, GEMs can be used to simulate disparate types of 'omics data, even though the explicit calculation of likelihoods may be intractable. Thus, the use of "simulation-based" Bayesian inference approaches is a promising route for informing GEM structure and parameters from data [198]. However, scaling Bayesian approaches up to deal with the large space of possible GEM reconstructions is an open, exciting and challenging research direction.

While this review has been entirely focused on uncertainty in GEM approaches, it is also important to remember that future efforts will need to creatively address major open questions on how to integrate metabolic models with other layers of biological complexity and their associated uncertainties. Several methods have been proposed to extend the basis of GEMs to include some other layers, such as metabolism and expression (ME) models that incorporate the processes of gene transcription and translation [199] or dynamic FBA that can simulate time courses of metabolic processes such as microbial growth curves [186, 200, 201], and can be extended to include multiple organisms and spatial structure [202–206]. Moving beyond the steady-state assumption, approaches based on kinetic models of metabolism can predict the concentrations of metabolites and fluxes through individual pathways. Although these models require a large number of kinetic parameters, beyond those required by GEMs, several methods exist for inferring these parameters and representing their uncertainty [207–209]. Finally, whole-cell modeling can be used to simultaneously model multiple processes in the cell and gain comprehensive insight into cellular physiology [210, 211]. However, considerable uncertainty in the many parameters required for kinetic and whole-cell modeling continues to limit their broad application [212, 213]. Thus, as new modeling approaches arise, it is likely that genome-scale metabolic modeling, which strikes a productive balance between scalability and scope with many successful applications [5–11], will continue to play a key role in the landscape of mechanistic modeling of biological systems. Further embracing uncertainty in this field is an exciting opportunity to continue to improve the application of this modeling framework.

Supplementary Information

The online version contains supplementary material available at <https://doi.org/10.1186/s13059-021-02289-z>.

Additional file 1. Review history.

Acknowledgements

We would like to acknowledge Alan Pacheco as well as all other members of the Segrè lab for useful discussion and feedback on the contents of this manuscript.

Peer review information

Andrew Cosgrove was the primary editor on this article and managed its editorial process and peer review in collaboration with the rest of the editorial team.

Review history

The review history is available as Additional File 1.

Authors' contributions

DB and SS wrote the original draft. All authors conceived of the manuscript content, edited the manuscript, and approved of the final manuscript.

Funding

This work was partially supported by the National Institute of Health (NIDCR R01DE024468, NIGMS R01GM121950, NIA UH2AG064704); the National Science Foundation (grants 1457695 and NSFOCE-BSF 1635070); the U.S. Department of Energy, Office of Science, Office of Biological & Environmental Research through the Microbial Community Analysis and Functional Evaluation in Soils SFA Program (m-CAFES) under contract number DE-AC02-05CH11231 to Lawrence Berkeley National Laboratory; the Human Frontiers Science Program (grant RGP0020/2016), the Boston University Interdisciplinary Biomedical Research Office, and by the Boston University training program in quantitative biology and physiology under Ruth L Kirschstein National Research Service Award T32GM008764 from the National Institute of General Medical Sciences. SS was funded by SINTEF, the Norwegian graduate research school in bioinformatics, biostatistics and systems biology (NORBIS) and by the INBioPharm project of the Centre for Digital Life Norway (Research Council of Norway grant no. 248885).

Competing interests

The authors declare that they have no competing interests.

Author details

¹Department of Biomedical Engineering and Biological Design Center, Boston University, Boston, MA, USA. ²Bioinformatics Program, Boston University, Boston, MA, USA. ³Department of Biotechnology and Food Science, NTNU - Norwegian University of Science and Technology, Trondheim, Norway. ⁴Department of Biotechnology and Nanomedicine, SINTEF Industry, Trondheim, Norway. ⁵K.G. Jebsen Center for Genetic Epidemiology, NTNU - Norwegian University of Science and Technology, Trondheim, Norway. ⁶Department of Biology and Department of Physics, Boston University, Boston, MA, USA.

Received: 22 July 2020 Accepted: 4 February 2021

Published online: 18 February 2021

References

- Maarleveld TR, Khandelwal RA, Olivier BG, Teusink B, Bruggeman FJ. Basic concepts and principles of stoichiometric modeling of metabolic networks. *Biotechnol J*. 2013;8:997–1008.
- Heiendt L, Arreckx S, Pfau T, Mendoza SN, Richelle A, Heinken A, et al. Creation and analysis of biochemical constraint-based models using the COBRA Toolbox v.3.0. *Nat Protoc*. 2019;14:639–702.
- Lewis NE, Nagarajan H, Palsson BO. Constraining the metabolic genotype–phenotype relationship using a phylogeny of in silico methods. *Nat Rev Microbiol*. 2012;10:291–305.
- Ebrahim A, Lerman JA, Palsson BO, Hyduke DR. COBRApy: Constraints-based reconstruction and analysis for python. *BMC Syst Biol*. 2013;7:74.
- Gu C, Kim GB, Kim WJ, Kim HU, Lee SY. Current status and applications of genome-scale metabolic models. *Genome Biol*. 2019;20:121.
- Cook DJ, Nielsen J. Genome-scale metabolic models applied to human health and disease. *WIREs Systems Biol Med*. 2017;9:e1393.
- Dunphy LJ, Papin JA. Biomedical applications of genome-scale metabolic network reconstructions of human pathogens. *Curr Opin Biotechnol*. 2018;51:70–9.
- Biggs MB, Medlock GL, Kolling GL, Papin JA. Metabolic network modeling of microbial communities. *WIREs Systems Biol Med*. 2015;7:317–34.
- Kim WJ, Kim HU, Lee SY. Current state and applications of microbial genome-scale metabolic models. *Current Opinion Systems Biol*. 2017;2:10–8.
- Zhang C, Hua Q. Applications of genome-scale metabolic models in biotechnology and systems medicine. *Front Physiol*. 2016;6.
- O'Brien EJ, Monk JM, Palsson BO. Using genome-scale models to predict biological capabilities. *Cell*. 2015;161:971–87.
- King ZA, Lu J, Dräger A, Miller P, Federowicz S, Lerman JA, et al. BiGG models: a platform for integrating, standardizing and sharing genome-scale models. *Nucleic Acids Res*. 2016;44:D515–22.
- Norsigian CJ, Pusarla N, McConn JL, Yurkovich JT, Dräger A, Palsson BO, et al. BiGG models 2020: multi-strain genome-scale models and expansion across the phylogenetic tree. *Nucleic Acids Res*. 2020;48:D402–6.
- Thiele I, Palsson BØ. A protocol for generating a high-quality genome-scale metabolic reconstruction. *Nat Protoc*. 2010;5:93–121.
- Henry CS, DeJongh M, Best AA, Frybarger PM, Linsay B, Stevens RL. High-throughput generation, optimization and analysis of genome-scale metabolic models. *Nature Biotechnology*. 2010;28:977–82.
- Karlsen E, Schulz C, Almaas E. Automated generation of genome-scale metabolic draft reconstructions based on KEGG. *BMC Bioinformatics*. 2018;19:467.
- Faria JP, Rocha M, Rocha I, Henry CS. Methods for automated genome-scale metabolic model reconstruction. *Biochem Soc Trans*. 2018;46:931–6.
- Seaver SMD, Liu F, Zhang Q, Jeffryes J, Faria JP, Edirisinghe JN, et al. The ModelSEED Biochemistry Database for the integration of metabolic annotations and the reconstruction, comparison and analysis of metabolic models for plants, fungi and microbes. *Nucleic Acids Res*. 2021;49:D575–88.
- Mendoza SN, Olivier BG, Molenaar D, Teusink B. A systematic assessment of current genome-scale metabolic reconstruction tools. *Genome Biol*. 2019;20:158.
- Arkin AP, Cottingham RW, Henry CS, Harris NL, Stevens RL, Maslov S, et al. KBase: the United States Department of Energy Systems Biology Knowledgebase. *Nat Biotechnol*. 2018;36:566–9.
- Monk J, Nogales J, Palsson BO. Optimizing genome-scale network reconstructions. *Nat Biotechnol*. 2014;32:447–52.
- Kirk PDW, Babbie AC, Stumpf MPH. Systems biology (un)certainities. *Science*. 2015;350:386–8.
- Orth JD, Thiele I, Palsson BØ. What is flux balance analysis? *Nature Biotechnol*. 2010;28:245–8.
- Feist AM, Herrgård MJ, Thiele I, Reed JL, Palsson BØ. Reconstruction of biochemical networks in microorganisms. *Nature Reviews Microbiol*. 2009;7:129–43.
- Wang L, Dash S, Ng CY, Maranas CD. A review of computational tools for design and reconstruction of metabolic pathways. *Synthetic Systems Biotechnol*. 2017;2:243–52.
- Labena AA, Gao Y-Z, Dong C, Hua H, Guo F-B. Metabolic pathway databases and model repositories. *Quant Biol*. 2018;6:30–9.
- Jing LS, Shah FFM, Mohamad MS, Hamran NL, Salleh AHM, Deris S, et al. Database and tools for metabolic network analysis. *Biotechnol Bioinform*. 2014;19:568–85.
- Tian W, Skolnick J. How well is enzyme function conserved as a function of pairwise sequence identity? *J Mol Biol*. 2003; 333:863–82.
- Schnoes AM, Brown SD, Dodevski I, Babbitt PC. Annotation error in public databases: misannotation of molecular function in enzyme superfamilies. *PLoS Computational Biol*. 2009;5:e1000605.
- Lobb B, Tremblay BJ-M, Moreno-Hagelsieb G, Doxey AC. An assessment of genome annotation coverage across the bacterial tree of life. *Microbial Genomics*. 2020;6:e000341.
- Ellens KW, Christian N, Singh C, Satagopam VP, May P, Linster CL. Confronting the catalytic dark matter encoded by sequenced genomes. *Nucleic Acids Res*. 2017;45:11495–514.
- Sorokina M, Stam M, Médigue C, Lespinet O, Vallenet D. Profiling the orphan enzymes. *Biol Direct*. 2014;9:10.

33. Griesemer M, Kimbrel JA, Zhou CE, Navid A, D'haeseleer P. Combining multiple functional annotation tools increases coverage of metabolic annotation. *BMC Genomics*. 2018;19:948.
34. Liberal R, Lisowska BK, Leak DJ, Pinney JW. PathwayBooster: a tool to support the curation of metabolic pathways. *BMC Bioinformatics*. 2015;16:86.
35. Ihmels J, Levy R, Barkai N. Principles of transcriptional control in the metabolic network of *Saccharomyces cerevisiae*. *Nat Biotechnol*. 2004;22:86–92.
36. Jacobs C, Lambourne L, Xia Y, Segrè D. Upon Accounting for the Impact of Isoenzyme Loss, Gene Deletion Costs Anticorrelate with Their Evolutionary Rates. *PLOS ONE*. 2017;12:e0170164.
37. Machado D, Herrgård MJ, Rocha I. Stoichiometric representation of gene–protein–reaction associations leverages constraint-based analysis from reaction to gene-level phenotype prediction. *PLoS Comput Biol*. 2016;12.
38. Agren R, Liu L, Shoaie S, Vongsangnak W, Nookaew I, Nielsen J. The RAVEN Toolbox and Its Use for Generating a Genome-scale Metabolic Model for *Penicillium chrysogenum*. *PLOS Computational Biol*. 2013;9:e1002980.
39. Wang H, Marcišauskas S, Sánchez BJ, Domenzain I, Hermansson D, Agren R, et al. RAVEN 2.0: a versatile toolbox for metabolic network reconstruction and a case study on *Streptomyces coelicolor*. *PLoS Comput Biol*. 2018;14:e1006541.
40. Aite M, Chevallier M, Frioux C, Trottier C, Got J, Cortés MP, et al. Traceability, reproducibility and wiki-exploration for “à-la-carte” reconstructions of genome-scale metabolic models. *PLoS Comput Biol*. 2018;14:e1006146.
41. Loira N, Zhukova A, Sherman DJ. Pantograph: a template-based method for genome-scale metabolic model reconstruction. *J Bioinforma Comput Biol*. 2015;13:1550006.
42. Hanemaaijer M, Olivier BG, Röling WFM, Bruggeman FJ, Teusink B. Model-based quantification of metabolic interactions from dynamic microbial-community data. *PLOS ONE*. 2017;12:e0173183.
43. Machado D, Andrejev S, Tramontano M, Patil KR. Fast automated reconstruction of genome-scale metabolic models for microbial species and communities. *Nucleic Acids Res*. 2018;46:7542–53.
44. Benedict MN, Mundy MB, Henry CS, Chia N, Price ND. Likelihood-based gene annotations for gap filling and quality assessment in genome-scale metabolic models. *PLoS Comput Biol*. 2014;10.
45. King B, Farrah T, Richards MA, Mundy M, Simeonidis E, Price ND. ProbAnnoWeb and ProbAnnoPy: probabilistic annotation and gap-filling of metabolic reconstructions. *Bioinformatics*. 2018;34:1594–6.
46. Pitkänen E, Jouhinen P, Hou J, Syed MF, Blomberg P, Kludas J, et al. Comparative genome-scale reconstruction of gapless metabolic networks for present and ancestral species. *PLoS Comput Biol*. 2014;10:e1003465.
47. Plata G, Fuhrer T, Hsiao T-L, Sauer U, Vitkup D. Global probabilistic annotation of metabolic networks enables enzyme discovery. *Nat Chem Biol*. 2012;8:848–54.
48. Ribeiro AJM, Holliday GL, Furnham N, Tyzack JD, Ferris K, Thornton JM. Mechanism and catalytic site Atlas (M-CSA): a database of enzyme reaction mechanisms and active sites. *Nucleic Acids Res*. 2018;46:D618–23.
49. Kautsar SA, Blin K, Shaw S, Navarro-Muñoz JC, Terlouw BR, van der Hoof JJJ, et al. MiBiG 2.0: a repository for biosynthetic gene clusters of known function. *Nucleic Acids Res*. 2020;48:D454–8.
50. Blin K, Shaw S, Steinke K, Villebro R, Ziemert N, Lee SY, et al. antiSMASH 5.0: updates to the secondary metabolite genome mining pipeline. *Nucleic Acids Res*. 2019;47:W81–7.
51. Elbourne LDH, Tetu SG, Hassan KA, Paulsen IT. TransportDB 2.0: a database for exploring membrane transporters in sequenced genomes from all domains of life. *Nucleic Acids Res*. 2017;45:D320–4.
52. Li H, Benedetto VA, Udvardi MK, Zhao PX. TransportTP: a two-phase classification approach for membrane transporter prediction and characterization. *BMC Bioinformatics*. 2009;10:418.
53. Price MN, Deutschbauer AM, Arkin AP. GapMind: Automated Annotation of Amino Acid Biosynthesis. *mSystems*. 2020;5.
54. Price MN, Wetmore KM, Waters RJ, Callaghan M, Ray J, Liu H, et al. Mutant phenotypes for thousands of bacterial genes of unknown function. *Nature*. 2018;557:503–9.
55. Fritzscheier CJ, Hartleb D, Szappanos B, Papp B, Lercher MJ. Erroneous energy-generating cycles in published genome scale metabolic networks: Identification and removal. *PLOS Computational Biol*. 2017;13:e1005494.
56. Ryu JY, Kim HU, Lee SY. Deep learning enables high-quality and high-throughput prediction of enzyme commission numbers. *PNAS*. 2019;116:13996–4001.
57. Klitgord N, Segrè D. The importance of compartmentalization in metabolic flux models: yeast as an ecosystem of organelles. *Genome Inform*. 2010;22:41–55.
58. Liu JK, O'Brien EJ, Lerman JA, Zengler K, Palsson BO, Feist AM. Reconstruction and modeling protein translocation and compartmentalization in *Escherichia coli* at the genome-scale. *BMC Syst Biol*. 2014;8:110.
59. Almagro Armenteros JJ, Sønderby CK, Sønderby SK, Nielsen H, Winther O. DeepLoc: prediction of protein subcellular localization using deep learning. *Bioinformatics*. 2017;33:3387–95.
60. Savojardo C, Martelli PL, Fariselli P, Proffiti G, Casadio R. BUSCA: an integrative web server to predict subcellular localization of proteins. *Nucleic Acids Res*. 2018;46:W459–66.
61. Price MN, Zane GM, Kuehl JV, Melnyk RA, Wall JD, Deutschbauer AM, et al. Filling gaps in bacterial amino acid biosynthesis pathways with high-throughput genetics. *PLoS Genetics*. 2018;14:e1007147.
62. Richards MA, Cassen V, Heavner BD, Ajami NE, Herrmann A, Simeonidis E, et al. MediaDB: a database of microbial growth conditions in defined media. *PLoS One*. 2014;9.
63. Oberhardt MA, Zarecki R, Gronow S, Lang E, Klenk H-P, Gophna U, et al. Harnessing the landscape of microbial culture media to predict new organism–media pairings. *Nature Communications*. 2015;6:1–14.
64. Aurich MK, Paglia G, Rolfsson Ó, Hrafnisdóttir M, Magnúsdóttir M, Stefaniak MM, et al. Prediction of intracellular metabolic states from extracellular metabolomic data. *Metabolomics*. 2015;11:603–19.
65. Zimmermann M, Kuehne A, Boshoff HJ, Barry CE, Zamboni N, Sauer U. Dynamic exometabolome analysis reveals active metabolic pathways in non-replicating mycobacteria. *Environ Microbiol*. 2015;17:4802–15.
66. Medlock GL, Carey MA, McDuffie DG, Mundy MB, Giallourou N, Swann JR, et al. Inferring Metabolic Mechanisms of Interaction within a Defined Gut Microbiota. *Cell Systems*. 2018;7:245–257.e7.
67. Venturilli OS, Carr AV, Fisher G, Hsu RH, Lau R, Bowen BP, et al. Deciphering microbial interactions in synthetic human gut microbiome communities. *Molecular Systems Biol*. 2018;14:e8157.
68. Øyås O, Borrell S, Trauner A, Zimmermann M, Feldmann J, Liphardt T, et al. Model-based integration of genomics and metabolomics reveals SNP functionality in mycobacterium tuberculosis. *PNAS*. 2020;117:8494–502.

69. Silva RR, Jourdan F, Salvanha DM, Letisse F, Jamin EL, Guidetti-Gonzalez S, et al. ProbMetab: an R package for Bayesian probabilistic annotation of LC-MS-based metabolomics. *Bioinformatics*. 2014;30:1336–7.
70. Marinos G, Kaleta C, Waschina S. Defining the nutritional input for genome-scale metabolic models: A roadmap. *PLOS ONE*. 2020;15:e0236890.
71. Edwards JS, Ibarra RU, Palsson BO. In silico predictions of *Escherichia coli* metabolic capabilities are consistent with experimental data. *Nature Biotechnol*. 2001;19:125–30.
72. Edwards JS, Ramakrishna R, Palsson BO. Characterizing the metabolic phenotype: a phenotype phase plane analysis. *Biotechnol Bioeng*. 2002;77:27–36.
73. Almaas E, Kovács B, Vicsek T, Oltvai ZN, Barabási A-L. Global organization of metabolic fluxes in the bacterium *Escherichia coli*. *Nature*. 2004;427:839–43.
74. Almaas E, Oltvai ZN, Barabási A-L. The Activity Reaction Core and Plasticity of Metabolic Networks. *PLOS Computational Biol*. 2005;1:e68.
75. Reed JL, Palsson BØ. Genome-scale in silico models of *E. coli* have multiple equivalent phenotypic states: assessment of correlated reaction subsets that comprise network states. *Genome Res*. 2004;14:1797–805.
76. Klier C. Use of an uncertainty analysis for genome-scale models as a prediction tool for microbial growth processes in subsurface environments. *Environ Sci Technol*. 2012;46:2790–8.
77. Ofaim S, Sulheim S, Almaas E, Sher DJ, Segrè D. Dynamic allocation of carbon storage and nutrient-dependent exudation in a revised genome-scale model of *Prochlorococcus*. *Front Genet Frontiers*. 2021;12
78. Klitgord N, Segrè D. Environments that Induce Synthetic Microbial Ecosystems. *PLOS Computational Biology*. 2010;6:e1001002.
79. Pacheco AR, Moel M, Segrè D. Costless metabolic secretions as drivers of interspecies interactions in microbial ecosystems. *Nature Communications*. 2019;10:1–12.
80. Bernstein DB, Dewhurst FE, Segrè D. Metabolic network percolation quantifies biosynthetic capabilities across the human oral microbiome. Shou W, Barkai N, Shou W, Quince C, editors. *eLife*. 2019;8:e39733.
81. Zarecki R, Oberhardt MA, Reshef L, Gophna U, Ruppin E. A novel nutritional predictor links microbial fastidiousness with lowered ubiquity, growth rate, and cooperativeness. *PLoS Comput Biol*. 2014;10
82. Andrade R, Wannagat M, Klein CC, Acuña V, Marchetti-Spaccamela A, Milreu PV, et al. Enumeration of minimal stoichiometric precursor sets in metabolic networks. *Algorithms for Molecular Biol*. 2016;11:25.
83. Seif Y, Choudhary KS, Hefner Y, Anand A, Yang L, Palsson BO. Metabolic and genetic basis for auxotrophies in gram-negative species. *PNAS*. 2020;117:6264–73.
84. Levy R, Borenstein E. Reverse ecology: from systems to environments and back. *Adv Exp Med Biol*. 2012;751:329–45.
85. Borenstein E, Kupiec M, Feldman MW, Ruppin E. Large-scale reconstruction and phylogenetic analysis of metabolic environments. *PNAS*. 2008;105:14482–7.
86. Feist AM, Palsson BO. The biomass objective function. *Curr Opin Microbiol*. 2010;13:344–9.
87. Xavier JC, Patil KR, Rocha I. Integration of biomass formulations of genome-scale metabolic models with experimental data reveals universally essential cofactors in prokaryotes. *Metab Eng*. 2017;39:200–8.
88. Yuan H, Cheung CYM, Hilbers PAJ, van Riel NAW. Flux balance analysis of plant metabolism: the effect of biomass composition and model structure on model predictions. *Front Plant Sci*. 2016;7
89. Volkmer B, Heinemann M. Condition-dependent cell volume and concentration of *Escherichia coli* to facilitate data conversion for systems biology modeling. *PLoS One*. 2011;6:e23126.
90. Schachter M, Maaløe O, Kjeldgaard NO. Dependency on medium and temperature of cell size and chemical composition during balanced growth of *Salmonella typhimurium*. *Microbiology*. 1958;19:592–606.
91. McKEE MJ, Knowles CO. Levels of protein, RNA, DNA, glycogen and lipid during growth and development of *Daphnia magna* Straus (Crustacea: Cladocera). *Freshw Biol*. 1987;18:341–51.
92. Chrzanowski TH, Grover JP. Element content of *Pseudomonas fluorescens* varies with growth rate and temperature: a replicated chemostat study addressing ecological stoichiometry. *Limnol Oceanogr*. 2008;53:1242–51.
93. Scott T, Cotner J, LaPara T. Variable stoichiometry and homeostatic regulation of bacterial biomass elemental composition. *Front Microbiol*. 2012;3
94. Carnicer M, Baumann K, Töplitz I, Sánchez-Ferrando F, Mattanovich D, Ferrer P, et al. Macromolecular and elemental composition analysis and extracellular metabolite balances of *Pichia pastoris* growing at different oxygen levels. *Microb Cell Factories*. 2009;8:65.
95. Cotner JB, Makino W, Biddanda BA. Temperature affects stoichiometry and biochemical composition of *Escherichia coli*. *Microb Ecol*. 2006;52:26–33.
96. Pramanik J, Keasling JD. Stoichiometric model of *Escherichia coli* metabolism: incorporation of growth-rate dependent biomass composition and mechanistic energy requirements. *Biotechnol Bioeng*. 1997;56:398–421.
97. Pramanik J, Keasling JD. Effect of *Escherichia coli* biomass composition on central metabolic fluxes predicted by a stoichiometric model. *Biotechnol Bioeng*. 1998;60:230–8.
98. Duarte NC, Herrgård MJ, Palsson BØ. Reconstruction and validation of *Saccharomyces cerevisiae* iND750, a fully compartmentalized genome-scale metabolic model. *Genome Res*. 2004;14:1298–309.
99. Nookaew I, Jewett MC, Meechai A, Thammarongtham C, Laoteng K, Cheevadhanarak S, et al. The genome-scale metabolic model iIN800 of *Saccharomyces cerevisiae* and its validation: a scaffold to query lipid metabolism. *BMC Syst Biol*. 2008;2:71.
100. Dikicioglu D, Kirdar B, Oliver SG. Biomass composition: the “elephant in the room” of metabolic modelling. *Metabolomics*. 2015;11:1690–701.
101. Puchalka J, Oberhardt MA, Godinho M, Bielecka A, Regenhardt D, Timmis KN, et al. Genome-scale reconstruction and analysis of the *Pseudomonas putida* KT2440 metabolic network facilitates applications in biotechnology. *PLoS Comput Biol*. 2008;4
102. Schulz C, Kumelj T, Karlsen E, Almaas E. Genome-scale metabolic modelling when changes in environmental conditions affect biomass composition. *bioRxiv*. 2020;2020.12.03.409565.
103. Beck AE, Hunt KA, Carlson RP. Measuring cellular biomass composition for computational biology applications. *Processes*. 2018;6:38.

104. Szélliová D, Ruckerbauer DE, Galleguillos SN, Petersen LB, Natter K, Hanscho M, et al. What CHO is made of: variations in the biomass composition of Chinese hamster ovary cell lines. *Metab Eng.* 2020;61:288–300.
105. Long CP, Antoniewicz MR. Quantifying biomass composition by gas chromatography/mass spectrometry. *Anal Chem.* 2014;86:9423–7.
106. Lachance J-C, Lloyd CJ, Monk JM, Yang L, Sastry AV, Seif Y, et al. BOFdat: generating biomass objective functions for genome-scale metabolic models from experimental data. *PLoS Comput Biol.* 2019;15:e1006971.
107. Mavrouniotis ML. Identification of qualitatively feasible metabolic pathways. *Artificial intelligence and molecular biology. USA: American Association for Artificial Intelligence.* 1993. p. 325–64.
108. Pan S, Reed JL. Advances in gap-filling genome-scale metabolic models and model-driven experiments lead to novel metabolic discoveries. *Curr Opin Biotechnol.* 2018;51:103–8.
109. Karp PD, Weaver D, Latendresse M. How accurate is automated gap filling of metabolic models? *BMC Syst Biol.* 2018;12:73.
110. Latendresse M, Karp PD. Evaluation of reaction gap-filling accuracy by randomization. *BMC Bioinformatics.* 2018;19:53.
111. Martyushenko N, Almaas E. ErrorTracer: an algorithm for identifying the origins of inconsistencies in genome-scale metabolic models. *Bioinformatics.* 2020;36:1644–6.
112. Green ML, Karp PD. A Bayesian method for identifying missing enzymes in predicted metabolic pathway databases. *BMC Bioinformatics.* 2004;5:76.
113. Dreyfuss JM, Zucker JD, Hood HM, Ocasio LR, Sachs MS, Galagan JE. Reconstruction and validation of a genome-scale metabolic model for the filamentous fungus *Neurospora crassa* using FARM. *PLoS Computational Biology.* 2013;9:e1003126.
114. Christian N, May P, Kempa S, Handorf T, Ebenhöf O. An integrative approach towards completing genome-scale metabolic networks. *Mol Biosyst.* 2009;5:1889–903.
115. Vitkin E, Shlomi T. MIRAGE: a functional genomics-based approach for metabolic network model reconstruction and its application to cyanobacteria networks. *Genome Biol.* 2012;13:R111.
116. Biggs MB, Papin JA. Managing uncertainty in metabolic network structure and improving predictions using EnsembleFBA. *PLoS Comput Biol.* 2017;13:e1005413.
117. Ponce-de-Leon M, Calle-Espinosa J, Perető J, Montero F. Consistency analysis of genome-scale models of bacterial metabolism: a metamodel approach. *PLOS ONE.* 2015;10:e0143626.
118. Krumholz EW, Libourel IGL. Sequence-based network completion reveals the integrality of missing reactions in metabolic networks. *J Biol Chem.* 2015;290:19197–19207.
119. Hadadi N, Hatzimanikatis V. Design of computational retrobiosynthesis tools for the design of de novo synthetic pathways. *Curr Opin Chem Biol.* 2015;28:99–104.
120. Prather KLJ, Martin CH. De novo biosynthetic pathways: rational design of microbial chemical factories. *Curr Opin Biotechnol.* 2008;19:468–74.
121. Hatzimanikatis V, Li C, Ionita JA, Henry CS, Jankowski MD, Broadbelt LJ. Exploring the diversity of complex metabolic networks. *Bioinformatics.* 2005;21:1603–9.
122. Jeffries JG, Colastani RL, Elbadawi-Sidhu M, Kind T, Niehaus TD, Broadbelt LJ, et al. MINe: open access databases of computationally predicted enzyme promiscuity products for untargeted metabolomics. *J Cheminformatics.* 2015;7:44.
123. Hafner J, MohammadiPeyhani H, Sveshnikova A, Scheidegger A, Hatzimanikatis V. Updated ATLAS of biochemistry with new metabolites and improved enzyme prediction power. *ACS Synth Biol.* 2020;9:1479–82.
124. Hadadi N, Hafner J, Shajkofci A, Zisaki A, Hatzimanikatis V. ATLAS of biochemistry: a repository of all possible biochemical reactions for synthetic biology and metabolic engineering studies. *ACS Synth Biol.* 2016;5:1155–66.
125. Price ND, Reed JL, Palsson BØ. Genome-scale models of microbial cells: evaluating the consequences of constraints. *Nat Rev Microbiol.* 2004;2:886–97.
126. Papoutsakis ET, Meyer CL. Equations and calculations of product yields and preferred pathways for butanediol and mixed-acid fermentations. *Biotechnol Bioeng.* 1985;27:50–66.
127. Fell DA, Small JR. Fat synthesis in adipose tissue. An examination of stoichiometric constraints. *Biochem J.* 1986;238:781–6.
128. Varma A, Palsson BO. Metabolic flux balancing: basic concepts, Scientific and Practical Use. *Nat Biotechnol.* 1994;12:994–8.
129. Edwards JS, Covert M, Palsson B. Metabolic modelling of microbes: the flux-balance approach. *Environ Microbiol.* 2002;4:133–40.
130. Schuetz R, Kuepfer L, Sauer U. Systematic evaluation of objective functions for predicting intracellular fluxes in *Escherichia coli*. *Molecular Systems Biol.* 2007;3:119.
131. Fong SS, Marciniak JY, Palsson BØ. Description and interpretation of adaptive evolution of *Escherichia coli* K-12 MG1655 by using a genome-scale in silico metabolic model. *J Bacteriol.* 2003;185:6400–8.
132. Gianchandani EP, Oberhardt MA, Burgard AP, Maranas CD, Papin JA. Predicting biological system objectives de novo from internal state measurements. *BMC Bioinformatics.* 2008;9:43.
133. Tenaillon O, Barrick JE, Ribeck N, Deatherage DE, Blanchard JL, Dasgupta A, et al. Tempo and mode of genome evolution in a 50,000-generation experiment. *Nature.* 2016;536:165–70.
134. Zhao Q, Stettner AI, Reznik E, Paschalidis IC, Segrè D. Mapping the landscape of metabolic goals of a cell. *Genome Biol.* 2016;17:109.
135. Harcombe WR, Delaney NF, Leiby N, Klitgord N, Marx CJ. The ability of flux balance analysis to predict evolution of central metabolism scales with the initial distance to the optimum. *PLoS Comput Biol.* 2013;9.
136. Segrè D, Vitkup D, Church GM. Analysis of optimality in natural and perturbed metabolic networks. *PNAS.* 2002;99:15112–7.
137. Wintermute EH, Lieberman TD, Silver PA. An objective function exploiting suboptimal solutions in metabolic networks. *BMC Syst Biol.* 2013;7:98.
138. Schuetz R, Zamboni N, Zampieri M, Heinemann M, Sauer U. Multidimensional optimality of microbial metabolism. *Science.* 2012;336:601–4.
139. Kitano H. Biological robustness. *Nat Rev Genet.* 2004;5:826–37.
140. Fischer E, Sauer U. Large-scale in vivo flux analysis shows rigidity and suboptimal performance of *Bacillus subtilis* metabolism. *Nat Genet.* 2005;37:636–40.
141. Mahadevan R, Schilling CH. The effects of alternate optimal solutions in constraint-based genome-scale metabolic models. *Metab Eng.* 2003;5:264–76.

142. Motamedian E, Naeimpoor F. LAMOS: a linear algorithm to identify the origin of multiple optimal flux distributions in metabolic networks. *Comput Chem Eng.* 2018;117:372–7.
143. Lee S, Phalakornkule C, Domach MM, Grossmann IE. Recursive MILP model for finding all the alternate optima in LP models for metabolic networks. *Comput Chem Eng.* 2000;24:711–6.
144. Maarleveld TR, Wortel MT, Olivier BG, Teusink B, Bruggeman FJ. Interplay between constraints, objectives, and optimality for genome-scale stoichiometric models. *PLOS Computational Biol.* 2015;11:e1004166.
145. Kelk SM, Olivier BG, Stougie L, Bruggeman FJ. Optimal flux spaces of genome-scale stoichiometric models are determined by a few subnetworks. *Sci Rep.* 2012;2:1–7.
146. Burgard AP, Nikolaev EV, Schilling CH, Maranas CD. Flux coupling analysis of genome-scale metabolic network reconstructions. *Genome Res.* 2004;14:301–12.
147. Schilling CH, Letscher D, Palsson BO. Theory for the systemic definition of metabolic pathways and their use in interpreting metabolic function from a pathway-oriented perspective. *J Theor Biol.* 2000;203:229–48.
148. Schuster S, Hilgetag C. On elementary flux modes in biochemical reaction systems at steady state. *J Biol Syst.* 1994;02:165–82.
149. Klamt S, Regensburger G, Gerstl MP, Jungreuthmayer C, Schuster S, Mahadevan R, et al. From elementary flux modes to elementary flux vectors: Metabolic pathway analysis with arbitrary linear flux constraints. *PLOS Computational Biol.* 2017;13:e1005409.
150. Ullah E, Yosafshahi M, Hassoun S. Towards scaling elementary flux mode computation. *Brief Bioinform.* 2020;21:1875–85.
151. Saa PA, Nielsen LK. LI-ACHRB: a scalable algorithm for sampling the feasible solution space of metabolic networks. *Bioinformatics.* 2016;32:2330–7.
152. Haraldsdóttir HS, Cousins B, Thiele I, Fleming RMT, Vempala S. CHRR: coordinate hit-and-run with rounding for uniform sampling of constraint-based models. *Bioinformatics.* 2017;33:1741–3.
153. Megchelenbrink W, Huynen M, Marchiori E optGpSampler: an improved tool for uniformly sampling the solution-space of genome-scale metabolic networks. *PLoS ONE* 2014;9:e86587.
154. Schellenberger J, Palsson BØ. Use of Randomized Sampling for Analysis of Metabolic Networks. *J Biol Chem.* 2009;284:5457–5461.
155. Wiback SJ, Famili I, Greenberg HJ, Palsson BØ. Monte Carlo sampling can be used to determine the size and shape of the steady-state flux space. *J Theor Biol.* 2004;228:437–47.
156. Price ND, Schellenberger J, Palsson BO. Uniform sampling of steady-state flux spaces: means to design experiments and to interpret enzymopathies. *Biophys J.* 2004;87:2172–86.
157. Bordel S, Agren R, Nielsen J. Sampling the solution space in genome-scale metabolic networks reveals transcriptional regulation in key enzymes. *PLoS Comput Biol.* 2010;6:e1000859.
158. Sulheim S, Kumpulij T, Dissel D van, Salehzadeh-Yazdi A, Du C, Wezel GP van, et al. Enzyme-constrained models and omics analysis of streptomyces coelicolor reveal metabolic changes that enhance heterologous production. *iScience.* 2020;23: 101525
159. Herrmann HA, Dyson BC, Vass L, Johnson GN, Schwartz J-M. Flux sampling is a powerful tool to study metabolism under changing environmental conditions. *NPJ Syst Biol Appl.* 2019;5:1–8.
160. Braunstein A, Muntoni AP, Pagnani A. An analytic approximation of the feasible space of metabolic networks. *Nat Commun.* 2017;8:1–9.
161. De Martino D, Andersson AM, Bergmiller T, Guet CC, Tkačik G. Statistical mechanics for metabolic networks during steady state growth. *Nat Commun.* 2018;9:1–9.
162. Fernandez-de-Cossio-Diaz J, Mulet R. maximum entropy and population heterogeneity in continuous cell cultures. *PLOS Computational Biology.* 2019;15:e1006823.
163. Jaynes ET. Information theory and statistical mechanics. *Phys Rev Am Physical Soc.* 1957;106:620–30.
164. Shore J, Johnson R. Axiomatic derivation of the principle of maximum entropy and the principle of minimum cross-entropy. *IEEE Trans Inf Theory.* 1980;26:26–37.
165. Heinonen M, Osmala M, Mannerström H, Wallenius J, Kaski S, Rousu J, et al. Bayesian metabolic flux analysis reveals intracellular flux couplings. *Bioinformatics.* 2019;35:i548–57.
166. Varma A, Boesch BW, Palsson BO. Stoichiometric interpretation of *Escherichia coli* glucose catabolism under various oxygenation rates. *Appl Environ Microbiol.* 1993;59:2465–73.
167. Pirt SJ, Hinshelwood CN. The maintenance energy of bacteria in growing cultures. *Proceedings of the Royal Society of London Series B Biological Sciences.* Royal Society. 1965;163:224–31.
168. Kempes CP, van Bodegom PM, Wolpert D, Libby E, Amend J, Hoehler T. Drivers of bacterial maintenance and minimal energy requirements. *Front Microbiol.* 2017;8
169. Opdam S, Richelle A, Kellman B, Li S, Zielinski DC, Lewis NE. A Systematic Evaluation of Methods for Tailoring Genome-Scale Metabolic Models. *Cels.* 2017;4:318–329.e6.
170. Goyal N, Padhiary M, Karimi IA, Zhou Z. Flux measurements and maintenance energy for carbon dioxide utilization by *Methanococcus maripaludis*. *Microb Cell Factories.* 2015;14:146.
171. Henry CS, Broadbelt LJ, Hatzimanikatis V. Thermodynamics-based metabolic flux analysis. *Biophys J.* 2007;92:1792–805.
172. Flamholz A, Noor E, Bar-Even A, Milo R. eQuilibrator—the biochemical thermodynamics calculator. *Nucleic Acids Res.* 2012;40:D770–5.
173. Noor E. Removing both Internal and Unrealistic Energy-Generating Cycles in Flux Balance Analysis. *arXiv:180304999 [q-bio]*. 2018;
174. Gerstl MP, Jungreuthmayer C, Zanghellini J. tEFMA: computing thermodynamically feasible elementary flux modes in metabolic networks. *Bioinformatics.* 2015;31:2232–4.
175. Noor E, Haraldsdóttir HS, Milo R, Fleming RMT. Consistent estimation of Gibbs energy using component contributions. *PLoS Comput Biol.* 2013;9:e1003098.
176. Jankowski MD, Henry CS, Broadbelt LJ, Hatzimanikatis V. Group contribution method for thermodynamic analysis of complex metabolic networks. *Biophys J.* 2008;95:1487–99.
177. Machado D, Herrgård M. Systematic evaluation of methods for integration of transcriptomic data into constraint-based models of metabolism. *PLOS Computational Biol.* 2014;10:e1003580.

178. Sánchez BJ, Zhang C, Nilsson A, Lahtvee P-J, Kerkhoven EJ, Nielsen J. Improving the phenotype predictions of a yeast genome-scale metabolic model by incorporating enzymatic constraints. *Molecular Systems Biol.* 2017;13:935.
179. Bekiaris PS, Klamt S. Automatic construction of metabolic models with enzyme constraints. *BMC Bioinformatics.* 2020;21:19.
180. Bathke J, Konzer A, Remes B, McIntosh M, Klug G. Comparative analyses of the variation of the transcriptome and proteome of *Rhodobacter sphaeroides* throughout growth. *BMC Genomics.* 2019;20:358.
181. Vogel C, Marcotte EM. Insights into the regulation of protein abundance from proteomic and transcriptomic analyses. *Nat Rev Genet.* 2012;13:227–32.
182. Lewis NE, Hixson KK, Conrad TM, Lerman JA, Charusanti P, Polpitiya AD, et al. Omic data from evolved *E. coli* are consistent with computed optimal growth from genome-scale models. *Molecular Systems Biol.* 2010;6:390.
183. Mori M, Hwa T, Martin OC, De Martino A, Marinari E. Constrained allocation flux balance analysis. *PLoS Comput Biol.* 2016;12:e1004913.
184. Niebel B, Leupold S, Heinemann M. An upper limit on Gibbs energy dissipation governs cellular metabolism. *Nat Metab.* 2019;1:125–32.
185. Moutinho TJ, Neubert BC, Jenior ML, Carey MA, Medlock GL, Kolling GL, et al. Functional anabolic network analysis of human-associated *Lactobacillus* strains. *bioRxiv.* 2019;746420.
186. Varma A, Palsson BO. Stoichiometric flux balance models quantitatively predict growth and metabolic by-product secretion in wild-type *Escherichia coli* W3110. *Appl Environ Microbiol.* 1994;60:3724–31.
187. Shamir M, Bar-On Y, Phillips R, Milo R. SnapShot timescales in cell biology. *Cell.* 2016;164:1302–1302.e1.
188. Zavanos MM, Julius AA. Robust flux balance analysis of metabolic networks. *Proceedings of the 2011 American Control Conference.* 2011. p. 2915–20.
189. MacGillivray M, Ko A, Gruber E, Sawyer M, Almaas E, Holder A. Robust analysis of fluxes in genome-scale metabolic pathways. *Sci Rep.* 2017;7:1–20.
190. Medlock GL, Moutinho TJ, Papin JA. Medusa: Software to build and analyze ensembles of genome-scale metabolic network reconstructions. *PLOS Computational Biol.* 2020;16:e1007847.
191. Stumpf MPH. Multi-model and network inference based on ensemble estimates: avoiding the madness of crowds. *J Royal Society Interface.* 2020;17:20200419.
192. Medlock GL, Papin JA. Guiding the refinement of biochemical knowledgebases with ensembles of metabolic networks and machine learning. *Cels.* 2020;10:109–119.e3.
193. Papin JA, Gabhann FM, Sauro HM, Nickerson D, Rampadarath A. Improving reproducibility in computational biology research. *PLOS Computational Biology.* 2020;16:e1007881.
194. Lieven C, Beber ME, Olivier BG, Bergmann FT, Ataman M, Babaei P, et al. MEMOTE for standardized genome-scale metabolic model testing. *Nature Biotechnol.* 2020;38:272–6.
195. Carey MA, Dräger A, Beber ME, Papin JA, Yurkovich JT. Community standards to facilitate development and address challenges in metabolic modeling. *Molecular Systems Biol.* 2020;16:e9235.
196. Noor E, Cherkaoui S, Sauer U. Biological insights through omics data integration. *Current Opinion Systems Biol.* 2019;15:39–47.
197. Ramon C, Gollub MG, Stelling J. Integrating –omics data into genome-scale metabolic network models: principles and challenges. *Essays Biochem.* 2018;62:563–74.
198. Cranmer K, Brehmer J, Louppe G. The frontier of simulation-based inference. *PNAS.* 2020;117:30055–62.
199. Lloyd CJ, Ebrahim A, Yang L, King ZA, Catoi E, O'Brien EJ, et al. COBRAME: A computational framework for genome-scale models of metabolism and gene expression. *PLOS Computational Biol.* 2018;14:e1006302.
200. Mahadevan R, Edwards JS, Doyle FJ. Dynamic flux balance analysis of diauxic growth in *Escherichia coli*. *Biophys J.* 2002; 83:1331–40.
201. Höffner K, Harwood SM, Barton PI. A reliable simulator for dynamic flux balance analysis. *Biotechnol Bioeng.* 2013;110: 792–802.
202. Harcombe WR, Riehl WJ, Dukovski I, Granger BR, Betts A, Lang AH, et al. Metabolic resource allocation in individual microbes determines ecosystem interactions and spatial dynamics. *Cell Rep.* 2014;7:1104–15.
203. Biggs MB, Papin JA. Novel multiscale modeling tool applied to *Pseudomonas aeruginosa* biofilm formation. *PLoS One.* 2013;8:e78011.
204. Bauer E, Zimmermann J, Baldini F, Thiele I, Kaleta C. BacArena: Individual-based metabolic modeling of heterogeneous microbes in complex communities. *PLOS Computational Biol.* 2017;13:e1005544.
205. Chen J, Gomez JA, Höffner K, Phalak P, Barton PI, Henson MA. Spatiotemporal modeling of microbial metabolism. *BMC Syst Biol.* 2016;10:21.
206. Borer B, Ataman M, Hatzimanikatis V, Or D. Modeling metabolic networks of individual bacterial agents in heterogeneous and dynamic soil habitats (IndiMeSH). *PLOS Computational Biol.* 2019;15:e1007127.
207. Andreatti S, Miskovic L, Hatzimanikatis V. iSCHRUNK – in Silico approach to characterization and reduction of uncertainty in the kinetic models of genome-scale metabolic networks. *Metab Eng.* 2016;33:158–68.
208. Miskovic L, Béal J, Moret M, Hatzimanikatis V. Uncertainty reduction in biochemical kinetic models: enforcing desired model properties. *PLOS Computational Biol.* 2019;15:e1007242.
209. PCS J, Strutz J, Broadbelt LJ, KEJ T, Bomble YJ. Bayesian inference of metabolic kinetics from genome-scale multiomics data. *PLOS Computational Biol.* 2019;15:e1007424.
210. Karr JR, Sanghvi JC, Macklin DN, Gutschow MV, Jacobs JM, Bolival B, et al. A whole-cell computational model predicts phenotype from genotype. *Cell Elsevier.* 2012;150:389–401.
211. Goldberg AP, Sziget B, Chew YH, Sekar JA, Roth YD, Karr JR. Emerging whole-cell modeling principles and methods. *Curr Opin Biotechnol.* 2018;51:97–102.
212. Babbie AC, Stumpf MPH. How to deal with parameters for whole-cell modelling. *J R Soc Interface.* 2017;14:20170237.
213. Saa PA, Nielsen LK. Formulation, construction and analysis of kinetic models of metabolism: a review of modelling frameworks. *Biotechnol Adv.* 2017;35:981–1003.

Publisher's Note

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

Paper 2

Predicting Strain Engineering Strategies Using iKS1317: A Genome-Scale Metabolic Model of *Streptomyces coelicolor*

Tjaša Kumelj, Snorre Sulheim, Alexander Wentzel and
Eivind Almaas.

Biotechnology journal, 14(4), 1800180 (2018).

Predicting Strain Engineering Strategies Using iKS1317: A Genome-Scale Metabolic Model of *Streptomyces coelicolor*

Tjaša Kumelj, Snorre Sulheim, Alexander Wentzel, and Eivind Almaas*

Streptomyces coelicolor is a model organism for the *Actinobacteria*, a phylum known to produce an extensive range of different bioactive compounds that include antibiotics currently used in the clinic. Biosynthetic gene clusters discovered in genomes of other *Actinobacteria* can be transferred to and expressed in *S. coelicolor*, making it a factory for heterologous production of secondary metabolites. Genome-scale metabolic reconstructions have successfully been used in several biotechnology applications to facilitate the over-production of target metabolites. Here, the authors present iKS1317, the most comprehensive and accurate reconstructed genome-scale metabolic model (GEM) for *S. coelicolor*. The model reconstruction is based on previous models, publicly available databases, and published literature and includes 1317 genes, 2119 reactions, and 1581 metabolites. It correctly predicts wild-type growth in 96.5% of the evaluated growth environments and gene knockout predictions in 78.4% when comparing with observed mutant growth phenotypes, with a total accuracy of 83.3%. However, using a minimal nutrient environment for the gene knockout predictions, iKS1317 has an accuracy of 87.1% in predicting mutant growth phenotypes. Furthermore, we used iKS1317 and existing strain design algorithms to suggest robust gene-knockout strategies to increase the production of acetyl-CoA. Since acetyl-CoA is the most important precursor for polyketide antibiotics, the suggested strategies may be implemented *in vivo* to improve the function of *S. coelicolor* as a heterologous expression host.

of new antibacterial drugs is a major threat to global health. One reason for the lack of novel drugs is that the traditional method of bioprospecting, involving cultivation and high-throughput screening, is no longer efficient, partly because of a high rate of rediscovery.^[1–3] A promising approach for discovery and production of novel bioactive metabolites is based on the heterologous expression of biosynthetic gene clusters in specialized expression host strains. One organism in which this already has been achieved is *Streptomyces coelicolor*.^[4–8]

The genus *Streptomyces* is one of the most important sources of bioactive, microbial metabolites. *S. coelicolor* is a model organism for this genus^[9] with a capacity to produce 31 different secondary metabolites,^[10] including four antibiotics: actinorhodin, calcium-dependent antibiotic, undecylprodigiosin and methylenomycin.^[11,12] Note that none of these four antibiotics are of medical relevance. Thus, it has a metabolic machinery capable of providing precursors for a large range of different classes of bioactive metabolites, which is a necessary feature of a host for heterologous expression of biosynthetic gene clusters and production of the encoded compounds. An improved *S. coelicolor* strain for heterologous expression has

already been developed by removing two plasmids naturally present in *S. coelicolor* A3(2) and four major biosynthetic gene clusters from the chromosome,^[13] resulting in a reduced metabolic and bioactive background. To further improve this


1. Introduction

The increasing resistance of pathogenic bacteria to antibiotics combined with a continuous low level of discovery and development

T. Kumelj, S. Sulheim, Prof. E. Almaas
Department of Biotechnology and Food Science
NTNU - Norwegian University of Science and Technology
Trondheim, Norway
E-mail: eivind.almaas@ntnu.no

S. Sulheim, Dr. A. Wentzel
SINTEF Industry
Department of Biotechnology and Nanomedicine
Trondheim, Norway

Prof. E. Almaas
K.G. Jebsen Center for Genetic Epidemiology
NTNU – Norwegian University of Science and Technology
Trondheim, Norway

 The ORCID identification number(s) for the author(s) of this article can be found under <https://doi.org/10.1002/biot.201800180>.

© 2018 The Authors. *Biotechnology Journal* Published by Wiley-VCH Verlag GmbH & Co. KGaA. This is an open access article under the terms of the Creative Commons Attribution License, which permits use, distribution and reproduction in any medium, provided the original work is properly cited.

DOI: 10.1002/biot.201800180

organism as an expression host, it is necessary to develop a more comprehensive understanding of the metabolism.

A genome-scale metabolic model (GEM) is a network representation of the metabolic capabilities of an organism, constructed from an annotated genome by using inferred or proven gene-protein-reaction relations, in addition to transport reactions, and an estimated biomass composition. A detailed reconstruction protocol is described in ref. [14]. The network of reactions and metabolites in a GEM can be mathematically represented by a stoichiometric matrix. Using a variety of constraint-based modeling approaches, this stoichiometric matrix serves as a core input to predict phenotypes of the organism subject to perturbations or its behavior in different growth environments. GEMs have been used successfully to direct strain engineering of organisms (e.g., see ref. [15] for a review).

Currently, there exist three genome-scale metabolic models for *S. coelicolor*; the iIB711,^[16] the iMA789,^[17] and the iMK1208,^[18] with the most recent being published in 2014. The iMA789 is an improved version of iIB711, and it includes a more comprehensive reconstruction of pathways for the production of antibiotics. The iMK1208 was reconstructed de novo based on annotations in StrepDB,^[19] Kyoto Encyclopedia of Genes and Genomes (KEGG),^[20] BioCyc,^[21] and TransportDB,^[22] and with updated biomass and ATP-maintenance reactions^[18].

Here, we present iKS1317, a more validated and comprehensive GEM of *S. coelicolor* based on the previous model iMK1208,^[18] appended and corrected with knowledge obtained from iMA789,^[17] KEGG,^[20] and BioCyc.^[21] Both reactions and metabolites are annotated with KEGG-identifiers when possible, and they are named according to guidelines and existing names in BiGG.^[23] The recent transposon mutagenesis study by Xu et al.^[24] enabled a thorough evaluation of the model's accuracy in predicting single gene knockout growth phenotypes. iKS1317 is written in the SBML format (level 3) and is compatible with both the COBRA Toolbox for Matlab and COBRApy.^[25,26]

Using iKS1317 as the basis for constraint-based optimization analyses, we suggest engineering strategies that may increase the heterologous production of polyketide antibiotics because of increased availability of the primary precursor acetyl-CoA. We predict the optimal yield of acetyl-CoA for *S. coelicolor* in response to single, double and triple reaction deletions in three different growth environments. Furthermore, we compare the results from OptKnock^[27] and Genetic Design through Local Search (GDLS),^[28] the two strain engineering methods used in this study.

2. Experimental Section

2.1. iKS1317 Model Reconstruction

The GEM presented here, iKS1317, is based on the iMK1208 model by Kim et al.^[18] An overview of the origin of reactions and metabolites in iKS1317 is given in Table 1. Of the 1859 reactions in iMK1208, 1852 were included in iKS1317 with no or minor changes. Four out of the seven removed reactions consisted of lumped reactions that we replaced by detailed, multi-reaction steps. One reaction was a duplicate, and two reactions in the actinorhodin pathway were mapped to reactions R09312 and R09313 in KEGG.^[20] A complete list of the removed reactions is provided in S1, Supporting information.

Based on the original biomass reaction in iMK1208, a second biomass reaction were constructed where the amino acids have been replaced by their respective tRNA-charged versions. The corresponding released tRNA molecules were added as products to balance the equation, and we adjusted the stoichiometric coefficients of ATP, ADP, water, protons, and phosphate to account for the energy consumed by the reactions charging the amino acids with tRNA molecules. The content of both biomass reactions is given in S2, Supporting information.

Metabolites in iMA789 and reactions in both iMA789 and iMK1208 were mapped to KEGG-identifiers,^[20] making it possible to directly compare the two models.^[17,18] The iMK1208 reconstruction is not based upon the iMA789, and we found 167 reactions in iMA789 not present in iMK1208 that we chose to include in iKS1317. The metabolites were mapped to KEGG-identifiers based on their name and chemical formula, and the reactions were mapped based on name, reactants, products, and co-factors. With the reactions annotated with KEGG-identifiers we could compare the content of iKS1317 with the list of reactions in KEGG associated to genes in the genome of *S. coelicolor* A3(2). This allowed us to find reactions in KEGG not already present in the model, and this investigation resulted in another 87 reactions appended to iKS1317. We assumed that the annotations in KEGG were correct if the reactions fitted well with the existing content of the model. If the KEGG-annotations were in contradiction to our existing model or involved a new pathway, they were further evaluated by using BioCyc, published literature, BRENDA, or Uniprot-SwissProt.^[21,29,30]

The metabolite formulas in iMA789 and KEGG are given in neutral (non-charged) form, while charged formulas are used in iMK1208. Most metabolites are charged in the cellular environment and this is also recommended in the 96-step protocol for model reconstruction by Thiele and Palsson.^[14] We calculated the charged chemical formula of the metabolites added from KEGG and iMA789 at pH 7 using eQuilibrator.^[31] Since not all chemical formulas could be calculated in this fashion, the formula for some metabolites were inferred by comparing neutral and charged formulas for similar metabolites.

eQuilibrator was also used to calculate the change in Gibbs free energy at standard conditions to infer reaction directionality in reactions from KEGG and iMA789. Because the concentration of each metabolite in the cell is unknown, we cannot accurately predict the change in Gibbs free energy of a reaction. Hence, we assumed most reactions to be reversible unless eQuilibrator predicted a large ($>30 \text{ kJ mol}^{-1}$)^[32,33] change in Gibbs free energy of the reaction.

Table 1. Origin of reactions and metabolites in the reconstructed metabolic network iKS1317.

	Reactions	Metabolites
iMK1208	1853	1435
iMA789	167	68
KEGG	87	69
BioCyc	12	9

A detailed overview of all reactions and their origin is found in S2, Supporting information.

By using the excellent review of the biosynthetic pathways in *S. coelicolor* by Challis,^[11] we added pathways for three secondary metabolites (geosmin, albaflavenon, methylenomycin) and extended the undecylprodigiosin pathway to include streptorubin. Most of these reactions were also described in BioCyc.^[21]

Fifty new transport reactions were added from iMA789, one of them transporting sucrose into the cytoplasm. The latter reaction enabled growth with sucrose as the sole carbon source, but according to Hodgson^[34] (as referred to by Borodina et al.^[16]) *S. coelicolor* is unable to catabolize sucrose. However, *S. coelicolor* is supposed to easily take up sucrose to balance the osmotic pressure (Bibb, 1985, as cited by Elibol, 1998).^[35,36] To avoid this erroneous in silico prediction, the uptake reaction rate for the sucrose-transport reaction was set to zero.

2.2. Validation of iKS1317

A listing of growth phenotyping data is available in the paper describing iMA711.^[16] Through additional literature review, we were able to identify growth data for 63 conditions for wild-type or mutant strains.^[16,34,37–44] The transposon mutagenesis data published by Xu et al.^[24] provide a valuable resource for model validation and improvement. This data set includes the growth phenotyping for 497 different gene knockout mutants, from which 365 are considered to be non-essential and 132 considered to be essential, unconditionally of the growth environment. The non-essential genes are confirmed by the ability of their knockout mutants to grow. The 132 unconditionally essential genes are identified by their lack of presence in any of the cultivated knockout mutants. Because the probability of having a gap in the genome in the transposon mutagenesis data increases with decreasing gap length, only gaps, and thus only genes, longer than 1.9 kb were included in the list of unconditionally essential genes.^[24] One hundred thirty-seven of these 497 different genes are present in iKS1317, of which 77 are non-essential and 60 are considered unconditionally essential genes. The transposon mutagenesis study was carried out on the *S. coelicolor* X737 mutant, from which all genes in the actinorhodin gene cluster are removed. We therefore removed these genes from our in silico model before computing the gene knockout growth phenotypes of iKS1317. In this study, we have assumed that the observations from the transposon mutagenesis study are absolutely correct, however this kind of large scale knockout experiments is difficult and can contain errors. A detailed overview of all tested growth conditions is found in S3, Supporting information.

We performed the comparison of in silico and in vivo growth in a binary fashion, classifying each condition as either *growth* or *no growth*. To compare growth rates in different environmental conditions we constrained the total uptake of carbon and nitrogen to 12.6 and 1.85 mmol (g dry weight h)⁻¹, respectively. This corresponds to the maximal experimentally observed uptake of glucose (2.1 mmol (g dry weight h)⁻¹)^[45] and the simulated corresponding ammonium uptake (1.85 mmol (g dry weight h)⁻¹). The specific uptake rates for each of the evaluated environments are given in S3, Supporting information. When we compared in silico predictions with experimental data, we needed to impose a lower threshold for the experimentally measurable growth rate. The measured growth rate for the maximal uptake rate of glucose is 0.128 h⁻¹.^[45]

corresponding to a doubling time of about 5.4 h. Consequently, we considered growth with a doubling time of more than 1 day as a reasonable lower threshold. This choice of threshold value only affected the growth on L-phenylalanine as carbon source, with a computationally predicted doubling time of 70 h.

For comparing our in silico predictions with growth data from the transposon mutagenesis study,^[24] we used a threshold of 50% of the in silico wild-type growth rate in the complex cultivation medium to decide the binary growth versus no growth test.^[46] This relatively large threshold was considered to be appropriate because of the methods used in a transposon mutagenesis study. We used the second biomass reaction where the amino acids in the primary biomass function are replaced by their respective tRNA charged versions to predict the knockout phenotypes. This enabled correct phenotype prediction for mutants where the knocked out genes are related to tRNA charging of amino acids.

The transposon mutants were sporulated and grown in complex media (SFM^[47] and YBP,^[48] respectively). This has direct implications for the modeling, since it is difficult to determine the carbon and nitrogen sources that actually were available and utilized by the organism. We therefore assumed that all carbon and nitrogen sources with an exchange reaction present in iKS1317 were available, a detailed list is given in S3, Supporting information. However, to shed light on some of the limitations of the use of the transposon mutagenesis data, we also included a comparison where the in silico growth of mutants was predicted with only glucose and ammonium available as the carbon and nitrogen source, respectively.

2.3. Suggesting Optimal Knockout Strategies with iKS1317

Two different strain design algorithms, OptKnock^[27] and GDLS,^[28] were used to predict genetic manipulations for target overproduction. The methods use constraint-based optimization to suggest reaction knockout (constraining the metabolic flux of a reaction to zero) strategies to gain targeted overproduction while optimizing internal (biomass yield) and external (product yield) cellular objectives. A direct consequence of these algorithms design is that overproduction of the target becomes an obligatory by-product of growth.^[27,28] These methods are accessible through the COBRA Toolbox v3.0^[25] in Matlab. The suggested reactions can be removed in vivo by knocking out one or more of the genes encoding the enzymes catalyzing the reaction.

The reaction knockout strategy was constrained with upper and lower bound on the ATP-maintenance reaction set to 2.65 mmol (g dry weight h)⁻¹^[18] and with a lower bound on the biomass growth rate of 0.05/h. The uptake rate of available carbon and nitrogen sources were set to 0.8 mmol (g dry weight h)⁻¹, except for the uptake of ammonium which was unlimited. We constrained OptKnock and GDLS to only allow knockout of enzymatic reactions with one or more associated genes, that is, all exchange reactions, 90 transport reactions, the biomass reactions, and the ATP-maintenance reaction were restricted from being removed. Additionally, 18 reactions related to oxidative phosphorylation were neither allowed to be knocked out (see S4, Supporting information, for details). The triple knockouts with OptKnock were computed on one node of an HPC platform with access to two Intel Xeon E5-2660 v3 CPUs.

We implemented the strain design algorithms with iKS1317 in three different growth environments: 1) a basic environment with glucose and ammonium; 2) a glucose-based environment enriched with nitrogen sources, that is, glutamate, nitrate, and ammonium; and 3) an environment enriched with carbon sources, that is, galactose, glycerol, mannitol, and with ammonium as the nitrogen source. An overview is given in Table 3A. By comparing the predicted strain engineering strategies in these three environments, we could evaluate the consistency of the suggested genetic modifications.

3. Results

The genome-scale metabolic model iKS1317 of *S. coelicolor*, contains 1317 genes (16% of the protein-coding genes in the genome),^[49–51] 2119 reactions and 1581 metabolites. An overview of the iKS1317 composition is given in Figure 1B. Both the model reactions and metabolites are now consistently annotated with KEGG IDs whenever possible. Reactions are annotated to 96 different pathways within 10 different subsystems, as defined by KEGG (Figure 1B).^[20,53] The additional reactions and genes relative to iMK1208 are mostly located in primary metabolism. However, we have also added pathways allowing the production of geosmin, albaflavenon, coenzyme F420, and methylenomycin, none of which are present in the previous model, iMK1208.^[18]

3.1. 83.3% Accuracy in Predicting Growth and Knockout Phenotypes

When we test the ability of our model to correctly predict experimental growth phenotypes, we find correct predictions in

96.5% (55/57) of the tested growth environments for the wild-type *S. coelicolor* (Table 2). Another approach for assessing the quality of a genome-scale reconstructed metabolic network, is to compare in vivo with in silico predictions of growth phenotypes for single-gene knockout mutants. This test is assessing the quality of the reconstruction of alternative metabolic pathways, and thus, is a complementary test to that of wild-type growth phenotyping. We found that iKS1317 predicts the correct knockout phenotype in 78.4% (120/153) of the compared conditions (Table 2). Eight of the correctly predicted knockout phenotypes are related to tRNA charging of the amino acids and would be false if the primary biomass reaction was used in the validation. In sum, iKS1317 has been evaluated in 210 different conditions and has an accuracy of 83.3% (175/210) in predicting growth and knockout phenotypes. The iMK1208 provides similar accuracy for predicting growth phenotypes (96.5%) and a 71.4% (105/147) accuracy for knockout mutants, resulting in an overall accuracy of 78.4% (160/204) (Figure 1A). Six of the 153 knockout phenotypes could not be evaluated by iMK1208 because the genes were not present in the model. A spreadsheet describing all growth comparisons is provided in S3, Supporting information.

3.2. OptKnock and GDLS Predicts Approximately 2-Fold Increase of Acetyl-CoA Production for Double-Knockout Mutants

Since the pyruvate dehydrogenase (PDH) reaction produces acetyl-CoA from pyruvate and coenzyme A (CoA), we selected it as the target reaction for the strain engineering analyses. The two strain engineering strategies, OptKnock^[27] and GDLS,^[28] predicted identical reaction knockouts as the optimal solution

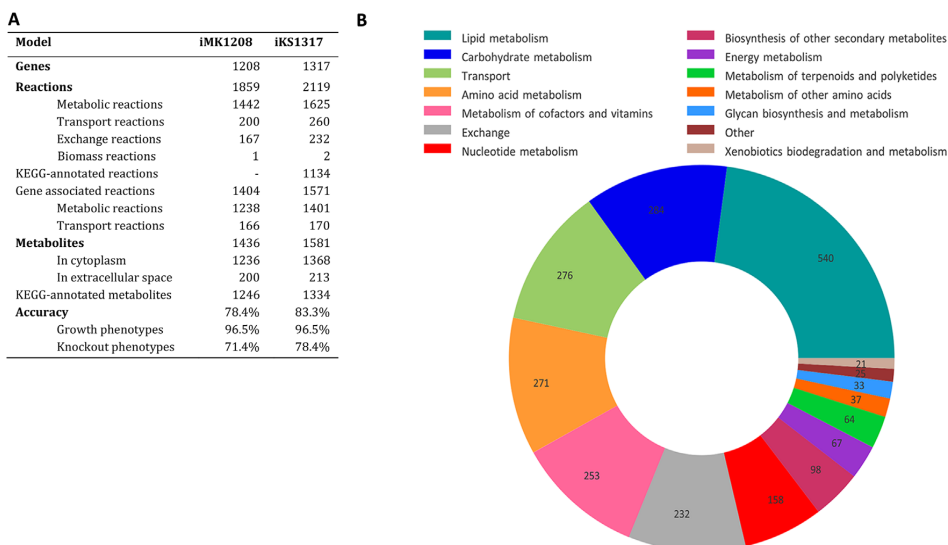


Figure 1. A) The table displays a side-by-side comparison of the two most recent genome-scale metabolic models for *S. coelicolor*. When assuming minimal cultivation medium with glucose and ammonium as the only carbon and nitrogen source, iKS1317 has a total accuracy of 90%. See Section 4 for details. B) The pie chart shows how reactions in iKS1317 are distributed over the main KEGG pathways, as well as exchange and transport reactions.

for single, double and triple knockouts in environments 1 and 3. For environment 2, both the single and double knockouts were identical, but GDLS provided a sub-optimal solution for the triple-knockout.

OptKnock also provides alternative solutions when several strategies are optimal (Strategy 7 and 8, Table 3B) and a pre-determined number of sub-optimal solutions (S4, Supporting information). The predicted production rates of acetyl-CoA through the PDH reaction are reported as the ratio of the rate in the knockout mutant to the rate in the wild-type in the same growth environment (Figure 2A, Table 3B). Absolute production rates, growth rates, associated genes, and sub-optimal solutions are provided in S4, Supporting information.

A numbering of the suggested strain engineering strategies and the reactions suggested for knockout are given in Table 3B. The name, KEGG ID and related pathway of these reactions are given in Table 3C. In the following section we refer to the strategies by the numbering and the reactions by their name.

The triple knockout in Environment 2 (Strategy 1) disrupt the synthesis of glutamate from alpha-ketoglutarate and provide the largest maximal relative production of acetyl-CoA (2.84). However, this strategy is not recommended for in vivo experiments because the minimal possible relative flux through PDH in this strategy is zero (Figure 2A). The double knockout in this environment (Strategy 2) provide a better solution: a large maximal relative rate of acetyl-CoA production (2.62) and a minimal relative equal the wild-type strain. In this strategy succinyl-CoA synthetase and glycine hydroxymethyltransferase is removed, and the effect on the flux distribution in the TCA cycle and central metabolism is displayed in Figure 2B. The reactions with major increase and decrease in flux is highlighted in red and blue, respectively. The removed succinyl-CoA synthetase is marked by a red X, and it is obvious that this has major impact on the flux distribution in the TCA cycle. We observe that the flux is rerouted out of the TCA cycle, into glyoxylate cycle to produce succinate and malate through isocitrate lyase and malate synthase. This increases the flux through the PDH reaction because malate is converted to pyruvate by malate dehydrogenase. Glycine hydroxymethyltransferase converts serine to glycine and is not shown in this figure, but the removal reroutes the synthesis of glycine through threonine. It is not obvious how this connects to the flux change in the TCA cycle, and this illustrates well why a computational

approach is necessary to identify the optimal engineering strategies.

Both the double and triple reaction knockouts in Environments 1 and 3 (Strategy 3–5) give similar changes in the flux distribution as Strategy 2 (Figure 2B). The overall pattern is that succinyl-CoA synthetase is knocked out along with either glutamate dehydrogenase or glycine hydroxymethyltransferase. The triple knockouts provide the largest predicted maximal production of acetyl-CoA in these environments (1.97 and 1.79 for Environments 1 and 3, respectively), but the small gain compared to Strategy 4 (1.80 and 1.71 for Environments 1 and 3, respectively) will probably not be worth the extra effort required to perform the additional knockout in vivo. The single reaction knockout strategies provide almost no increase in the production of acetyl-CoA (Figure 2A).

4. Discussion

This work presents iKS1317, an updated genome-scale metabolic network for *S. coelicolor*, providing more accurate predictions than any previous genome-scale reconstruction for this organism (Table 1). The iKS1317 network is also more comprehensive, more thoroughly annotated, and better validated than previous models.

Many of the discrepancies between in vivo growth and in silico predictions present in previous metabolic reconstructions of *S. coelicolor* are related to the degradation and biosynthesis of branched-chain amino-acids, and they have been resolved in iKS1317. It is known that the *S. coelicolor* Δvdh (SCO4089) knockout mutant is unable to grow with L-valine, L-leucine, or L-isoleucine as the sole carbon source,^[43] an observation not supported by iMK1208.^[18] However, by only changing the L-valine (R01214), L-leucine (R01090), and L-isoleucine (R02199) transaminase reactions from reversible to irreversible, these reactions are prevented from participating in the degradation of branched-chain amino acids, and these experimental results are recovered in iKS1317. In contradiction, the estimated changes in Gibbs free energy at 1 mM concentration and standard conditions are 3.2 ± 6.9 , -1.2 ± 3.2 , and -4.7 ± 6.9 kJ mol⁻¹ for R01214, R01090, and R02199, respectively,^[31] not indicating that the reactions are irreversible in any direction. However, different metabolite concentrations and the efficiency of upstream and downstream reactions have a major impact on these values.

Second, the $\Delta msdA$ (SCO2726) knockout mutant is incapable of growth in vivo, with L-valine as the sole carbon source. The gene *msdA* encodes for the enzyme methylmalonate-semialdehyde dehydrogenase which catalyze the reactions methylmalonate-semialdehyde: NAD⁺ oxidoreductase (R00935) and 3-oxopropanoate:NAD⁺ oxidoreductase (R00705). Removing the reaction methylmalonate semialdehyde: NAD⁺ oxidoreductase disrupts the primary degradation pathway of L-valine, but according to the metabolic reconstruction L-valine can also be degraded through the pathways for biosynthesis and degradation of L-leucine. This connection is possible because of the enzyme catalysing the reaction 2-isopropylmalate synthase (R01213). We have in iKS1317 introduced a redox coupling (acetyl-CoA/CoA) between 3-oxopropanoate: NAD⁺ oxidoreductase (R00705) and

Table 2. This table details the result of the comparison between in silico predictions and in vivo observations.

	Growth environments		Knockout mutants		
	In silico	Growth	No growth	Growth	
In vivo	Growth	TP: 51	FN: 0	TP: 78	FN: 6
	No growth	FP: 2	TN: 4	FP: 27	TN: 42

The left part displays the growth phenotypes for 57 different growth environments and the right part display growth phenotypes for 153 different knockout mutants. The two false positive predictions for the growth environments are with glutamine and aspartate as the sole carbon source. The 27 false predictions for the knockout mutants are examined in the Discussion section. The following abbreviations are used: true positives (TP), false positives (FP), false negatives (FN), and true negatives (TN).

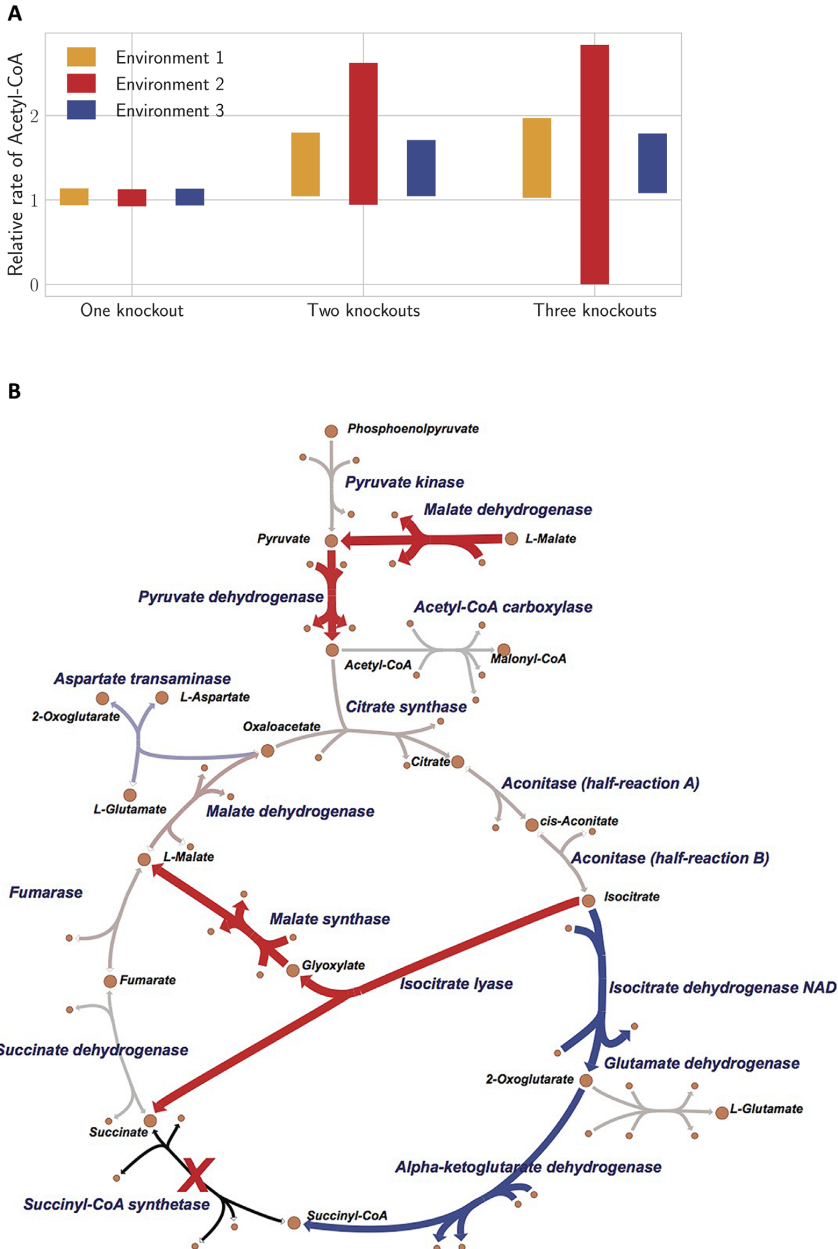


Figure 2. A) Predicted relative production rate ranges of the primary antibiotic precursor acetyl-CoA through the pyruvate dehydrogenase (PDH) reaction for single, double, and triple reaction knockout mutants in the three different growth environments (Table 3A). Each bar spans between the minimum and maximum computed rates, relative to the rate of the PDH reaction for the wild-type in the same growth environment. OptKnock^[27] and GDLS^[28] provided identical predictions except for the triple knockout in environment 3, where GDLS returned sub-optimal solutions. B) The predicted change in flux distribution when succinyl-CoA synthetase (marked by red X) and glycine hydroxymethyltransferase (not in figure) are knocked out (Strategy 2, Table 3B). The flux is rerouted out of the TCA cycle and into the glyoxylate cycle to produce succinate and malate. This increases the flux through malate dehydrogenase and pyruvate dehydrogenase. The map is drawn using Escher and display only the TCA cycle and related reactions in iKS1317.^[52] The reactions with major increase and decrease in flux are highlighted in red and blue, respectively.

Table 3. An overview of the strain engineering environments and results.

(A) Environments			
Environment no.	Carbon sources	Nitrogen sources	
1	Glucose	Ammonium	
2	Glucose	Glutamate ^a , nitrate, ammonium	
3	Galactose, glycerol, mannitol	Ammonium	

(B) Strain engineering predictions			
Strategy no.	Environment no.	Reactions	Max production
1	2	R00114, R00093, R00248	2.84
2	2	R00405, R00945	2.62
3	1	R00405, R00658, R00945	1.97
4	1	R00405, R00248	1.80
5	3	R00405, R00248, R00955	1.79
4	3	R00405, R00248	1.71
6	2	R00844	1.05
7	1	R04780	1.04
8	1	R01070	1.04
9	3	R00248	1.03

(C) KEGG ID, name and pathway of suggested reactions		
KEGG ID	Reaction name	Pathway
R00093	Glutamate synthase	Glutamate metabolism
R00114	Glutamate synthase	Glutamate metabolism
R00248	Glutamate dehydrogenase	Glutamate metabolism
R00405	Succinyl-CoA synthetase	TCA-cycle
R00658	Enolase	Glycolysis/gluconeogenesis
R00844	Glycerol-3-phosphate dehydrogenase	Glycerophospholipid metabolism
R00945	Glycine hydroxymethyltransferase	Glycine, serine and threonine metabolism
R00955	Uridyl transferase	Galactose metabolism
R01070	Fructose-bisphosphate aldolase	Glycolysis
R04780	Fructose 1,6-bisphosphatase	Gluconeogenesis

Table 3A display the carbon and nitrogen sources of the three environments. Table 3B display the predicted optimal strain engineering strategies: The maximal production is the relative rate of the pyruvate dehydrogenase reaction (PDH) with respect to the maximal rate of this reaction for the wild-type in the same growth environment. Table 3C display the KEGG ID, name and pathway of each of the reactions in the suggested strategies. The absolute minimal and maximal production rates, growth rates, associated genes, and sub-optimal solutions are provided in S4, Supporting information. ^aGlutamate is a source of both carbon and nitrogen.

2-isopropylmalate synthase (R01213) which blocks the latter reaction for the $\Delta msdA$ knockout mutant and solves this discrepancy between in vivo observation and in silico prediction. A detailed description is given in S5, Supporting information.

While our analysis and manual curation of the metabolic network reconstruction has removed the two before mentioned discrepancies, we are still left with an incorrect growth prediction of the $\Delta msdA$ (SCO2726) mutant with propionate as the sole carbon source. In contrast to in vivo experiments,^[37] the iKS1317 predicts no growth, since the introduced coupling also disrupts the synthesis of L-leucine, which is an essential amino acid. Further research is required to fully understand the regulatory mechanisms involved in the synthesis and degradation of the branched-chain amino acids in *S. coelicolor*.

In iKS1317, we propose a new, possible pathway for the biosynthesis of L-isoleucine: four intermediate steps catabolizing pyruvate to 2-oxobutanoate. According to KEGG, *S. coelicolor* is missing the initial reaction of this pathway, R-citramalate synthase. However, upon conducting a protein BLAST (BLASTP) search,^[54] we uncovered a sequence similarity of 47% (E-value: 2e-168) between SCO5529 and the *cimA* gene in *Geobacter sulfurreducens*, where this pathway has been experimentally validated.^[55] The same reaction has also been identified in vivo in *Cyanobacteria*, and it has been suggested that this pathway may be present in other organisms.^[56] The output from our BLASTP search is available in S6, Supporting information.

We have used published transposon mutagenesis data^[24] to validate and correct our model predictions for essential genes. In contradiction to the iMK1208 model predictions,^[18] the

transposon study indicates that SCO5626 is an essential gene encoding the ATP/UMP phosphotransferase. According to the gene annotations in iMK1208, the *cmk* gene (SCO1760) also encodes for this enzyme. However, the enzyme encoded by *cmk* is highly specific for the ATP/CMP phosphotransferase in prokaryotes.^[57] Consequently, we removed it from this gene-reaction rule in iKS1317. Another apparently essential gene is SCO3894, encoding a transmembrane protein involved in the murein biosynthesis. This differs in KEGG and iMK1208, which suggest that the enzyme encoded by SCO2709 is an isozyme of the enzyme encoded by SCO3894. A BLASTP search of the *murJ* gene in *E. coli*, where the function of the encoded protein is shown,^[58] show high similarity with both SCO2709 (E-value 3e-13) and SCO3894 (E-value 1e-20), and we have therefore kept SCO2709 and SCO3894 as isogenes in iMK1208. The output from our BLASTP search is available in S7, Supporting information.

The disagreements between our model predictions and in vivo observed wild-type growth phenotypes are associated with the utilization of glutamine and aspartate as carbon sources.^[34] These erroneous phenotypes are also predicted in iMA789 and iMK1208, and it has previously been suggested that this lack of in vivo growth is caused by regulatory effects.^[16] It is surprising that *S. coelicolor* is unable to grow on L-aspartate, because it is observed in vivo that it can utilize L-asparagine as a carbon source,^[34] which is then degraded to L-aspartate. Thus, by only taking the topology of the metabolic network into account, the organism should be able to grow on L-aspartate as well.

Contrary to our predictions, *S. coelicolor* is unable to grow with glutamine as the carbon source (^[34] as cited in ref. ^[16]). However, glutamine can be used as a nitrogen source, providing ammonium through conversion to glutamate by glutaminase intracellularly. Glutamate is further decarboxylated into γ -aminobutanoate upon uptake.^[59] Borodina et al.^[16] suggested that the intracellular glutamate provided by glutaminase cannot be further degraded, explaining why glutamine can function as a nitrogen source but not a carbon source. On the other hand, a *AgnA* (SCO2198) mutant, lacking glutamine synthase, can utilize glutamine as a carbon source.^[39] According to optimization theory it is not possible to increase the metabolic repertoire of a GEM through a gene knockout, since a gene knockout reduces the solution space.^[60] This observation is in support of the suggestion that a regulatory effect is the cause for the observed lack of utilization of glutamine as a carbon source in vivo.

The growth predictions for knockout mutants may seem less accurate (78.4% than the growth phenotype predictions (96.5%). The erroneous predictions are mostly false positives, that is, iKS1317 predicts growth for knockout mutants that do not grow in vivo. This indicates that the model is too flexible and contains optional pathways or isogenes when an in vivo essential gene is knocked out. As is customary for constraint-based modeling, unless explicit knowledge is present, one does not account for possible differences in reaction rates between alternative enzymes or pathways.^[60] Thus, a pathway providing a perfect replacement in silico may actually be too slow to support detectable growth in vivo. While it is possible to use enzyme kinetics to limit the reaction rates in a GEM, it is difficult to acquire reliable values.^[61]

We determined 27 false positive predictions for the different knockout mutants, that is, 27 genes that are observed as essential in vivo but not in silico. Fourteen of the false positive predictions can be traced back to uncertainty of the nutrient environment, more specifically the available carbon and nitrogen sources present in the complex growth medium in which the transposon mutagenesis mutants were cultivated.^[24] One example is the knockout of the gene *ddl* (SCO5560), encoding the D-alanine-D-alanine ligase, which is observed to be an essential gene in the transposon mutagenesis study, in contradiction to the iKS1317 predictions. However, by simply removing D-alanyl-D-alanine from the growth medium in silico, iKS1317 correctly predicts no growth for the Δ *ddl* mutant. These 14 false positives are changed to the true negative category if we assume a minimal medium with glucose and ammonium as the sole carbon and nitrogen sources, respectively. With this assumption, iKS1317 has an accuracy of 87.1% (134/153) for predicting knockout phenotypes, resulting in a total accuracy of 90% (189/210). This demonstrates a challenge with the typical use of transposon mutagenesis data for model curation and validation.

Another limiting factor is the lower bound of 1.9 kb on the length of the genes identified as essential reactions. Of the 1317 genes present in iKS1317, only 129 (9.8%) are longer than 1.9 kb. A more extensive transposon mutagenesis study with a lower threshold would increase value of the data because it could enable the evaluation of a larger number of essential genes.

We observe that many of the genes annotated to their respective enzymes, and thus reactions, are inferred from homology with similar genes in different organisms. It is in most cases not certain that these genes actually encode for the same enzymes, potentially leading to erroneous model predictions. Another possibility which will provide false-positive results is that a gene that is present in the iKS1317 model may not be expressed in vivo. There exist several methods for using transcriptomics (see Ref. ^[62] for a comparison) to restrict the model solution space to only include genes expressed in the chosen conditions, and such data may improve model predictions.

By using OptKnock^[27] and GDLS^[28] we have predicted optimal single, double, and triple reaction-knockout strategies to increase the production of acetyl-CoA through PDH. The GDLS algorithm use heuristics to perform a local search, and it is not guaranteed to find the optimal solution.^[28] With the GDLS algorithm the optimal solution was found for all single and double reaction knockouts, and in two of the three environments for the triple knockouts. While OptKnock always find the optimal solution, the CPU-time for a triple reaction knockout using OptKnock and iKS1317 is about 160 h on a HPC-platform, compared to a few minutes with GDLS on an average laptop. Thus, for more than three knockouts with iKS1317, OptKnock becomes practically infeasible when all possible reactions are considered.

Production of acetyl-CoA was selected as a target for the strain engineering algorithms because it is the most important precursor for biosynthesis of polyketides. Increasing the precursor pool has previously provided increased secondary metabolite production in *S. coelicolor*^[63,64] and similar strains.^[65–67] Increasing the precursor pool can be combined with overexpression of the biosynthetic gene cluster encoding the

pathway producing the target compound to further improve the likelihood of increased production.^[68]

Several precise regulatory mechanisms are involved in the biosynthesis of antibiotics in vivo. Among these mechanisms, carbon-source interplay appears to be one of the main factors controlling secondary metabolism.^[69,70] Therefore, reaction deletion strategies for overproduction of polyketide antibiotics in silico were suggested in various growth environments (Table 3A), that included different nitrogen and especially carbon sources. Both Strategy 2 and 4 seem like good suggestions for in vivo strain optimization: Strategy 2 provides the largest increase in acetyl-CoA production of these two, but Strategy 4 is more robust to different growth environments.

When comparing the suggested strain engineering strategies with experimental data, we find good agreement with the results from Huang et al.^[71] that found increased production of FK506 in a *ΔgdhA* *Streptomyces tsukubaensis* knockout mutant. FK506 is a combined polyketide synthase and non-ribosomal peptide with acetyl-CoA as one of the main precursors, and *ΔgdhA* encodes for glutamate dehydrogenase (R00248), one of the suggested knockouts in Strategy 1, 4, 5, and 9 (Table 3B).

However, when taking the transposon mutagenesis data into account, some of the suggested strategies are in contradiction to observed knockout mutants: the *sucC* (SCO4808) and *sucD* (SCO4809) genes encoding the succinyl-CoA synthetase complex are classified as essential genes.^[24] Succinyl-CoA synthetase (R00405) is one of the two knocked out reactions in both Strategy 2 and 4 (Table 3B). According to iKS1317 and protein BLAST SCO4808 and SCO4809 are not essential because SCO6585 and SCO6586 are isogenes encoding the same enzyme complex with E-values of 2e-145 and 1e-156, respectively (S8, Supporting information). Additionally, iKS1317 predicts less than 2% reduction in growth rate if succinyl-CoA synthetase is knocked out. Possible reasons for this discrepancy include: 1) The genes SCO6585 and SCO6586 may not be expressed in the cultivation media used in the transposon mutagenesis experiment^[24] and 2) iKS1317 is too flexible and predicts an efficient flux rerouting when succinyl-CoA synthetase is removed which does not occur in vivo. Consequently, our computational in-depth analysis of iKS1317 serves as an example of the systems-biology science iteration paradigm, by producing further hypothesis that need experimental follow up.

Abbreviations

GDLS, genetic design through local search; GEM, genome-scale metabolic model; KEGG, kyoto encyclopedia of genes and genomes.

Supporting information

Supporting Information is available from the Wiley Online Library or from the author.

Acknowledgement

T.K. and S.S. contributed equally to this work. This research is a part of the INBioPharm project of the Centre for Digital Life Norway (project no. 248885), funded by The Research Council of Norway.

Conflict of Interest

The authors declare no commercial or financial conflict of interest.

Keywords

constraint-based optimization, genome-scale model, *Streptomyces coelicolor*

Received: April 8, 2018
Revised: November 15, 2018
Published online: January 16, 2019

- [1] C. L. Ventola, *Pharm. Ther.* **2015**, *40*, 277.
- [2] C. L. Ventola, *Pharm. Ther.* **2015**, *40*, 344.
- [3] D. J. Payne, M. N. Gwynn, D. J. Holmes, D. L. Pompliano, *Nat. Rev. Drug Discov.* **2007**, *6*, 29.
- [4] H. Ikeda, K. Shin-ya, S. Omura, *J. Ind. Microbiol. Biotechnol.* **2014**, *41*, 233.
- [5] J. Yin, M. Hoffmann, X. Bian, Q. Tu, F. Yan, L. Xia, X. Ding, A. F. Stewart, R. Müller, J. Fu, *Sci. Rep.* **2015**, *5*, 15081.
- [6] B. Bonet, R. Teufel, M. Crüsemann, N. Ziemert, B. S. Moore, *J. Nat. Prod.* **2014**, *78*, 539.
- [7] K. Yamanaka, K. A. Reynolds, R. D. Kersten, K. S. Ryan, D. J. Gonzalez, V. Nizet, P. C. Dorrestein, B. S. Moore, *Proc. Natl. Acad. Sci.* **2014**, *111*, 1957.
- [8] J. P. Gomez-Escribano, M. J. Bibb, *J. Ind. Microbiol. Biotechnol.* **2014**, *41*, 425.
- [9] J. Berdy, *J. Antibiot.* **2005**, *58*, 1.
- [10] M. Nett, H. Ikeda, B. S. Moore, *Nat. Prod. Rep.* **2009**, *26*, 1362.
- [11] G. L. Challis, *J. Ind. Microbiol. Biotechnol.* **2014**, *41*, 219.
- [12] D. A. Hopwood, *Streptomyces in Nature and Medicine: The Antibiotic Makers*. Oxford University Press, Oxford, UK **2007**.
- [13] J. P. Gomez-Escribano, M. J. Bibb, *Microb. Biotechnol.* **2011**, *4*, 207.
- [14] I. Thiele, B. Ø. Palsson, *Nat. Protoc.* **2010**, *5*, 93.
- [15] E. Simeonidis, N. D. Price, *J. Ind. Microb. Biotechnol.* **2015**, *42*, 327.
- [16] I. Borodina, P. Krabben, J. Nielsen, *Genome Res.* **2005**, *15*, 820.
- [17] M. T. Alam, M. E. Merlo, D. A. Hodgson, E. M. Wellington, E. Takano, R. Breitling, T. S. C. (stream), *BMC Genomics* **2010**, *11*, 202.
- [18] M. Kim, J. Sang Yi, J. Kim, J.-N. Kim, M. W. Kim, B.-G. Kim, *Biotechnol. J.* **2014**, *9*, 1185.
- [19] <http://strepdb.streptomyces.org.uk>
- [20] M. Kanehisa, S. Goto, *Nucleic Acids Res.* **2000**, *28*, 27.
- [21] R. Caspi, R. Billington, L. Ferrer, H. Foerster, C. A. Fulcher, I. M. Keseler, A. Kothari, M. Krummenacker, M. Latendresse, L. A. Mueller, Q. Ong, S. Paley, P. Subhraveti, D. S. Weaver, P. D. Karp, *Nucleic Acids Res.* **2015**, *44*, D471.
- [22] L. D. Elbourne, S. G. Tetu, K. A. Hassan, I. T. Paulsen, *Nucleic Acids Res.* **2016**, *45*, D320.
- [23] Z. A. King, J. Lu, A. Dräger, P. Miller, S. Federowicz, J. A. Lerman, A. Ebrahim, B. Ø. Palsson, N. E. Lewis, *Nucleic Acids Res.* **2016**, *44*, D515.
- [24] Z. Xu, Y. Wang, K. F. Chater, H.-Y. Ou, H. H. Xu, Z. Deng, M. Tao, *Appl. Environ. Microbiol.* **2017**, *83*, e02889.
- [25] L. Heirendt, S. Arreckx, T. Pfau, S. N. Mendoza, A. Richelle, A. Heinken, H. S. Haraldsdottir, S. M. Keating, V. Vlasov, J. Wachowiak, S. Magnusdottir, C. Yu Ng, G. Preciat, A. Zagare, S. H. Chan, M. K. Aurich, C. M. Clancy, J. Modamio, J. T. Sauls, A. Noronha, A. Bordbar, B. Cousins, A. M. Assal, CEL Diana, M. B. Guenila, Apaolaza, A. Kostromins, H. M. Le, S. Y. Ma, Ding, L. V. Valcarcel, L. Wang, J. T. Yurkovich, P. T. Young, H. S. Assal, P. Lemmer El, W. A. Bryant, F. J. Aragoien Artacho, F. J. Planes,

- E. Stalidzans, V. S. Maass, Alejandro, M. Hucka, M. A. Saunders, C. D. Maranas, N. E. Lewis, T. Sauter, B. Ø. Palsson, I. Thiele, R. M. Fleming, *Nat. Protoc.* **2011**, 6, 1290.
- [26] A. Ebrahim, J. A. Lerman, B. Ø. Palsson, D. R. Hyduke, *BMC Syst. Biol.* **2013**, 7, 74.
- [27] A. P. Buargard, P. Pharkya, C. D. Maranas, *Biotechnol. Bioeng.* **2003**, 84, 647.
- [28] D. S. Lun, G. Rockwell, N. J. Guido, M. Baym, J. A. Kelner, B. Berger, Galagan, G. M. J. E. Church, *Mol. Syst. Biol.* **2009**, 5, 1.
- [29] M. Scheer, A. Grote, A. Chang, I. Schomburg, C. Munaretto, M. Rother, C. Söhngen, M. Stelzer, J. Thiele, D. Schomburg, *Nucleic Acids Res.* **2010**, 39, D670.
- [30] U. Consortium, *Nucleic Acids Res.* **2018**, 46, 2699.
- [31] A. Flamholz, E. Noor, A. Bar-Even, R. Milo, *Nucleic Acids Res.* **2011**, 40, D770.
- [32] A. Bar-Even, A. Flamholz, E. Noor, R. Milo, *Biochim. Biophys. Acta (BBA)-Bioenerg.* **2012**, 1817, 1646.
- [33] N. Elad, A. Bar-Even, A. Flamholz, Y. Lubling, D. Davidi, R. Milo, *Bioinformatics* **2012**, 2012, 2037.
- [34] D. Hodgson, Carbohydrate utilization in *Streptomyces coelicolor* A3 (2), *Ph.D. Thesis*, University of East Anglia **1980**.
- [35] M. Elibol, F. Mavituna, *Process Biochem.* **1998**, 33, 307.
- [36] M. Bibb, D. Hopwood, K. Chater, T. Kieser, C. Bruton, H. Kieser, D. Lydiate, C. Smith, J. Ward, H. Schrepf, *Genetic Manipulation of Streptomyces: A Laboratory Manual*. John Innes Foundation, Norwich, UK **1986**.
- [37] Y.-X. Zhang, L. Tang, C. R. Hutchinson, *J. Bacteriol.* **1996**, 178, 490.
- [38] G. Van Wezel, J. White, M. Bibb, P. Postma, *Mol. Gen. Genet. MGG* **1997**, 254, 604.
- [39] D. Fink, D. Falke, W. Wohlleben, A. Engels, *Microbiology* **1999**, 145, 2313.
- [40] F. Barona-Gómez, D. A. Hodgson, *EMBO Rep.* **2003**, 4, 296.
- [41] F. Flett, J. Platt, J. Cullum, *J. Basic Microbiol.* **1987**, 27, 1.
- [42] D.-J. Hu, D. Hood, R. Heidstra, D. Hodgson, *Mol. Microbiol.* **1999**, 32, 869.
- [43] L. Tang, C. Hutchinson, *J. Bacteriol.* **1993**, 175, 4176.
- [44] M. Fischer, C. Schmidt, D. Falke, R. G. Sawers, *Res. Microbiol.* **2012**, 163, 340.
- [45] K. Melzoch, M. J. T. de Mattos, O. M. Neijssel, *Biotechnol. Bioeng.* **1997**, 54, 577.
- [46] B. Papp, C. Pal, L. D. Hurst, *Nature* **2004**, 429, 661.
- [47] G. Hobbs, C. M. Frazer, D. C. Gardner, J. A. Cullum, S. G. Oliver, *Appl. Microbiol. Biotechnol.* **1989**, 31, 272.
- [48] X. Ou, B. Zhang, L. Zhang, G. Zhao, X. Ding, *Appl. Environ. Microbiol.* **2009**, 75, 2158.
- [49] S. D. Bentley, K. F. Chater, A.-M. Cerdeño-Tárraga, G. L. Challis, N. Thomson, K. D. James, D. E. Harris, M. A. Quail, H. Kieser, D. Harper, A. Bateman, S. Brown, G. Chandra, C. W. Chen, M. Collins, A. Cronin, A. Fraser, A. Goble, J. Hidalgo, T. Hornsby, S. Howarth, C.-H. Huang, T. Kieser, L. Larke, L. Murphy, K. Oliver, S. O'Neil, E. Rabinowitz, M.-A. Rajandream, K. Rutherford, S. Rutter, K. Seeger, D. Saunders, S. Sharp, R. Squares, S. Squares, K. Taylor, T. Warren, A. Wietzorrek, J. Woodward, B. G. Barrell, J. Parkhill, D. A. Hopwood, *Nature* **2002**, 417, 141.
- [50] S. D. Bentley, S. Brown, L. D. Murphy, D. E. Harris, M. A. Quail, J. Parkhill, B. G. Barrell, J. R. McCormick, R. I. Santamaria, R. Losick, M. Yamasaki, H. Kinashi, C. W. Chen, G. Chandra, D. Jakimowicz, H. M. Kieser, T. Kieser, K. F. Chater, *Mol. Microbiol.* **2004**, 51, 1615.
- [51] I. Haug, A. Weissenborn, D. Brolle, S. Bentley, T. Kieser, J. Altenbuchner, *Microbiology* **2003**, 149, 505.
- [52] Z. A. King, A. Dräger, A. Ebrahim, N. Sonnenschein, N. E. Lewis, B. Ø. Palsson, *PLoS Comput. Biol.* **2015**, 11, e1004321.
- [53] This use of subsystems comes from Pathway maps for Metabolism in KEGG (<http://www.genome.jp/kegg/pathway.html>).
- [54] S. F. Altschul, W. Gish, W. Miller, E. W. Myers, D. J. Lipman, *J. Mol. Biol.* **1990**, 215, 403.
- [55] C. Rizzo, S. J. Van Dien, A. Orloff, D. R. Lovley, M. V. Coppi, *J. Bacteriol.* **2008**, 190, 2266.
- [56] B. Wu, B. Zhang, X. Feng, J. R. Rubens, R. Huang, L. M. Hicks, H. B. Pakrasi, Y. J. Tang, *Microbiology* **2010**, 156, 596.
- [57] T. Bertrand, P. Briozzo, L. Assairi, A. Ofiteru, N. Bucurenci, H. Munier-Lehmann, B. Golinelli Pimpaneau, O. Bârzau, A.-M. Gilles, *J. Mol. Biol.* **2002**, 315, 1099.
- [58] L.-T. Sham, E. K. Butler, M. D. Lebar, D. Kahne, T. G. Bernhardt, N. Ruiz, *Science* **2014**, 345, 220.
- [59] L. Inbar, A. Lapidot, *J. Bacteriol.* **1991**, 173, 7790.
- [60] B. Ø. Palsson, *Systems Biology: Constraint-Based Reconstruction and Analysis*. Cold Spring Harbor Laboratory Press (CSH Press), Cold Spring Harbor, New York, USA **2015**.
- [61] R. Adadi, B. Volkmer, R. Milo, M. Heinemann, T. Shlomi, *PLoS Comput. Biol.* **2012**, 8, e1002575.
- [62] D. Machado, M. Herrgard, *PLoS Comput. Biol.* **2014**, 10, e1003580.
- [63] Y.-G. Ryu, M. J. Butler, K. F. Chater, K. J. Lee, *Appl. Environ. Microbiol.* **2006**, 72, 7132.
- [64] Y. J. Kim, J. Y. Song, M. H. Moon, C. P. Smith, S.-K. Hong, Y. K. Chang, *Appl. Microbiol. Biotechnol.* **2007**, 76, 1119.
- [65] S. Mo, Y.-H. Ban, J. W. Park, Y. J. Yoo, Y. J. Yoon, *J. Ind. Microbiol. Biotechnol.* **2009**, 36, 1473.
- [66] D. Zabala, A. F. Braña, A. B. Flórez, J. A. Salas, C. Méndez, *Metab. Eng.* **2013**, 20, 187.
- [67] H. L. Robertsen, T. Weber, H. U. Kim, S. Y. Lee, *Biotechnol. J.* **2018**, 13, 1700465.
- [68] W. S. Jung, E. Kim, Y. J. Yoo, Y. H. Ban, E. J. Kim, Y. J. Yoon, *Appl. Microbiol. Biotechnol.* **2014**, 98, 3701.
- [69] A. Wentzel, P. Bruheim, A. Øverby, M. Ø. Jakobsen, H. Sletta, W. A. Omara, D. A. Hodgson, T. E. Ellingsen, *BMC Syst. Biol.* **2012**, 6, 59.
- [70] S. Sanchez, A. Chávez, A. Forero, Y. García-Huante, A. Romero, M. Sánchez, D. Rocha, B. Sánchez, M. Avalos, S. Guzmán-Trampe, R. Rodríguez-Sanoja, E. Langley, B. Ruiz, *J. Antibiot.* **2010**, 63, 442.
- [71] D. Huang, S. Li, M. Xia, J. Wen, X. Jia, *Microb. Cell Fact.* **2013**, 12, 52.

Paper 3

Automatic reconstruction of metabolic pathways from identified biosynthetic gene clusters

Snorre Sulheim, Fredrik A Fossheim, Alexander Wentzel and Eivind Almaas.

BMC Bioinformatics 22, 81 (2021).

RESEARCH ARTICLE

Open Access



Automatic reconstruction of metabolic pathways from identified biosynthetic gene clusters

Snorre Sulheim^{1,2*}, Fredrik A. Fosshem¹, Alexander Wentzel² and Eivind Almaas^{1,3}

*Correspondence:
snorre.sulheim@sintef.no

¹ Department of Biotechnology and Food Science, NTNU - Norwegian University of Science and Technology, Sem Sælands vei 8, 7034 Trondheim, Norway
Full list of author information is available at the end of the article

Abstract

Background: A wide range of bioactive compounds is produced by enzymes and enzymatic complexes encoded in biosynthetic gene clusters (BGCs). These BGCs can be identified and functionally annotated based on their DNA sequence. Candidates for further research and development may be prioritized based on properties such as their functional annotation, (dis)similarity to known BGCs, and bioactivity assays. Production of the target compound in the native strain is often not achievable, rendering heterologous expression in an optimized host strain as a promising alternative. Genome-scale metabolic models are frequently used to guide strain development, but large-scale incorporation and testing of heterologous production of complex natural products in this framework is hampered by the amount of manual work required to translate annotated BGCs to metabolic pathways. To this end, we have developed a pipeline for an automated reconstruction of BGC associated metabolic pathways responsible for the synthesis of non-ribosomal peptides and polyketides, two of the dominant classes of bioactive compounds.

Results: The developed pipeline correctly predicts 72.8% of the metabolic reactions in a detailed evaluation of 8 different BGCs comprising 228 functional domains. By introducing the reconstructed pathways into a genome-scale metabolic model we demonstrate that this level of accuracy is sufficient to make reliable *in silico* predictions with respect to production rate and gene knockout targets. Furthermore, we apply the pipeline to a large BGC database and reconstruct 943 metabolic pathways. We identify 17 enzymatic reactions using high-throughput assessment of potential knockout targets for increasing the production of any of the associated compounds. However, the targets only provide a relative increase of up to 6% compared to wild-type production rates.

Conclusion: With this pipeline we pave the way for an extended use of genome-scale metabolic models in strain design of heterologous expression hosts. In this context, we identified generic knockout targets for the increased production of heterologous compounds. However, as the predicted increase is minor for any of the single-reaction knockout targets, these results indicate that more sophisticated strain-engineering strategies are necessary for the development of efficient BGC expression hosts.



Keywords: Biosynthetic gene clusters, Genome-scale metabolic model, AntiSMASH, Polyketide synthases, Natural products, Heterologous expression, Non-ribosomal peptide synthetases

Background

Natural products provide an immense source of bioactive small molecules of medical and agricultural importance [1–3]. The biosynthesis of these small-molecule bioactive compounds is usually governed by genes that are clustered in physical close proximity on the genome in fungal [4] or bacterial species [5], commonly known as biosynthetic gene clusters (BGCs). The revolution in sequencing technology has enabled access to complete genome sequences for an increasing number of bacteria and fungi. Mining of these genomes has revealed a vast abundance of BGCs, many more than the number of bioactive compounds observed *in vitro* [6, 7], suggesting that many BGCs are not expressed or that their respective compounds are not produced at detectable amounts in laboratory conditions. The activation of these silent BGCs may lead to the discovery of many novel bio-pharmaceuticals [8].

One promising avenue towards exploration of the bioactive potential of these silent BGCs is heterologous expression in host strains that are engineered to achieve maximal production of the encoded natural products [9, 10]. With current software [11] it is possible to quickly mine a genome for BGCs and retrieve information about the class, location, and functional domains of every gene in each cluster [12]. One may further prioritize BGC candidates for heterologous expression based on this information, (dis)similarity to known BGCs, bioactivity assays and mass spectrometry profiles of produced compounds, and subsequently transfer the selected BGCs to a chosen host strain using available genetic tools [13, 14]. However, the cloning and transfer of BGCs can be time-consuming and difficult depending on the genetic tools available for the native and the heterologous host strains, as well as the size of the BGC in question [15]. Additionally, it is not clear which host strain or which genetic modifications will maximize the yield of the secondary metabolite synthesized through the metabolic pathway catalyzed by the enzymes, or enzyme complexes, encoded by the heterologously expressed BGC [16, 17].

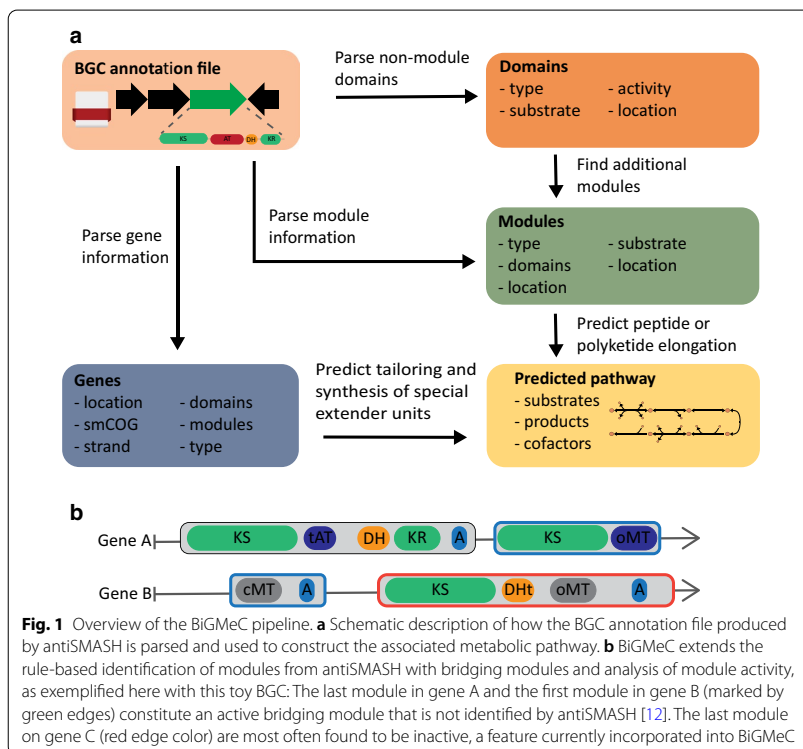
Genome-scale metabolic models (GEMs) can predict the consequence of genetic modifications [18] and are routinely used to guide strain design for a wide range of purposes [19]. However, this approach has still not gained traction in guiding strain-engineering efforts to increase the heterologous production of complex natural compounds, despite a number of available GEMs for *Actinobacteria* [20], a phylum known for an extremely diverse secondary metabolism responsible for about two-thirds of all known antibiotics in use today [21]. Previous efforts are limited to maximization of native secondary metabolites [22–24] or precursor pools [25]. One reason for the lack of computational efforts leveraging GEMs to assess heterologous production from BGCs is the significant amount of work required to map out the associated metabolic pathway, although most of the required information is contained in the output from software used to identify and annotate BGCs, such as antiSMASH [12]. In this work, we address this hurdle by developing a pipeline that parses the output obtained from antiSMASH and constructs

the corresponding metabolic-synthesis pathway, thereby making BGCs available for constraint-based analysis and strain engineering guided by GEMs.

We have chosen to focus on non-ribosomal peptide synthetases (NRPSs) and two types of polyketide synthases (PKSs), namely type 1 PKSs and trans-AT PKSs. These BGC classes are of particular interest because of their vast abundance [26, 27] and great prospect to become novel biopharmaceuticals [28, 29]. For an exhaustive description of NRPS and PKS biosynthesis, we refer the reader to a range of excellent reviews [27, 30–33], but we provide the brief summary required as a context for the later description of the pipeline and results. Both NRPS, and type 1 and trans-AT PKS biosynthesis are performed by multidomain enzyme complexes that create a polymer from amino acid or acyl-CoA building blocks, respectively. The chain elongation is performed by well-defined modules that makes it tractable to predict the biosynthetic pathways producing the associated compounds from the annotated sequence data, but the presence of iterative modules can complicate predictions [34–36]. An active chain elongating module in an NRPS cluster requires at least three functional domains: a condensation (C) domain, an adenylation (A) domain and a peptidyl carrier (PCP) domain. The A domain activates a specific amino acid (or in some cases a carboxylic acid) and facilitates the attachment of the amino acid to the PCP domain, while the C domain catalyzes the formation of peptide bonds required to elongate the peptide. In addition to these three domains, NRPS modules can replace the C domain by a Cy domain performing condensation and heterocyclisation or additionally contain a methyltransferase (MT) and/or an epimerase (E) domain. The load module initiating biosynthesis usually lacks the C domain, while the terminating module contains either a thioesterase (TE) or a thioester reductase (TR) domain.

Similar to NRPSs, chain elongating modules of PKSs rely on three functional domains: an acyltransferase (AT) domain that recognizes a specific extender unit and attaches it to the acyl carrier (ACP) domain. The third domain, ketosynthase (KS) catalyzes the Claisen condensation required to extend the polyketide chain. A standard PKS load module contains only the AT and ACP domain, and a TE or TR domain is required for the release of the polyketide chain by the final PKS module. PKS modules can also feature the reducing domains ketoreductase (KR), dehydratase (DH) and enoylreductase (ER), and different combinations of functional domains yield a large variety of molecular transformations, in particular for the trans-AT PKSs [32]. These trans-AT PKSs not only differ from normal (*cis*) modular PKSs by having a larger module diversity and deviations from canonical rules, but they are also recognized by freestanding AT domains that perform the chain elongation [32]. The diversity of PKS and NRPS natural products is further extended by hybrid variants containing both NRPS and PKS domains and modules.

We acknowledge that experimental analyses of the final and intermediate products, as well as enzyme activity assays, are required to fully unravel the details of the metabolic pathways associated with a BGC. However, for the chosen classes of BGCs (NRPS, type 1 PKS, and trans-AT PKS), we hypothesize that the information acquired from genome mining is sufficient to make *in silico* predictions that are biologically relevant. After assembling and evaluating the accuracy of the new pipeline presented in this work, we demonstrate its value towards high-throughput assessment of BGCs by reconstructing



the metabolic pathways for 943 of the BGCs currently in MIBiG [37]. Furthermore, we predict the optimal single reaction inactivation (by gene knockout) strain-engineering strategy for natural product synthesis based on each BGC when introduced into a genome-scale metabolic model of *Streptomyces coelicolor*, a model organism among the *Actinobacteria* and a popular heterologous BGC expression host [15, 38].

Results

We have developed the Biosynthetic Gene cluster Metabolic pathway Construction (BiGMeC) pipeline that leverages antiSMASH results to create the metabolic pathway corresponding to a PKS or NRPS biosynthetic gene cluster (Fig. 1a). The pipeline details each enzymatic reaction of the metabolic pathway, including redox cofactors and energy demand. The results are stored in a format that is easily introduced into a GEM using popular tools for constraint-based reconstruction and analysis, such as cobrapy [39] or COBRA Toolbox [40].

The hallmarks of PKS- and NRPS-genes are adjacent functional domains that in total make up one or several modules that initiate, extend or cleave off the polyketide or peptide product, respectively [30, 32, 33]. The output from antiSMASH comprises information about these modules and their functional domains, and occasionally also the specific extender unit or chemical transformation associated with each functional domain [12]. The BiGMeC pipeline not only parses this information, but uses well-reasoned heuristics

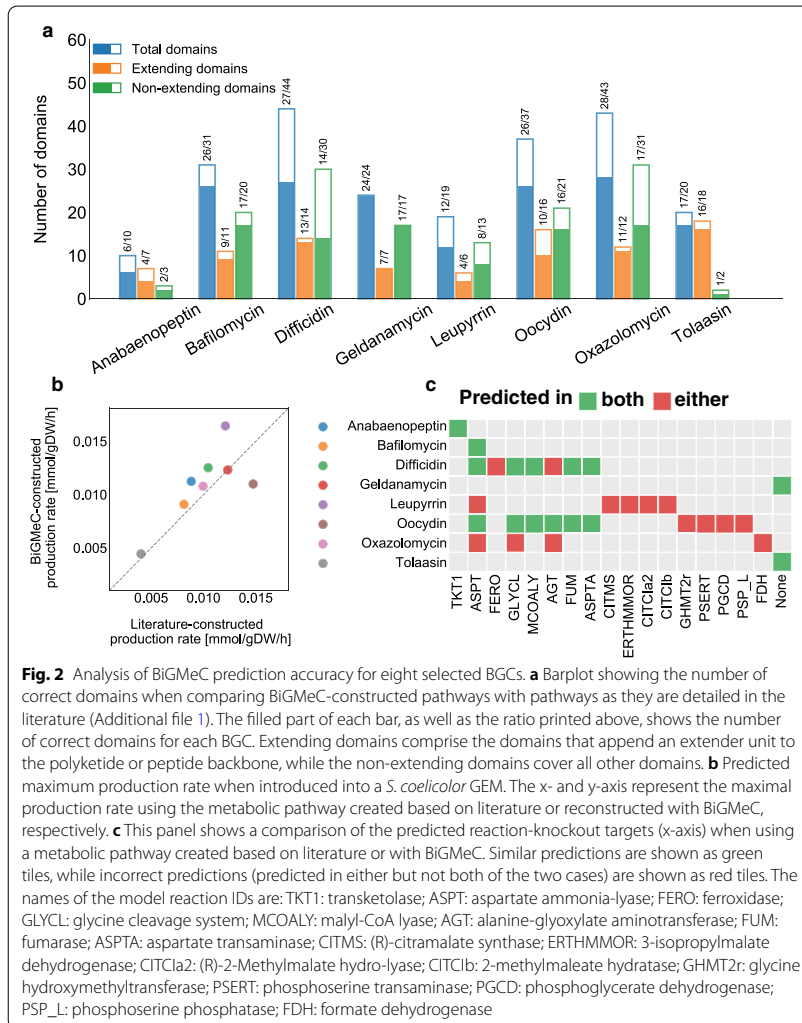


Fig. 2 Analysis of BiGMeC prediction accuracy for eight selected BGCs. **a** Barplot showing the number of correct domains when comparing BiGMeC-constructed pathways with pathways as they are detailed in the literature (Additional file 1). The filled part of each bar, as well as the ratio printed above, shows the number of correct domains for each BGC. Extending domains comprise the domains that append an extender unit to the polyketide or peptide backbone, while the non-extending domains cover all other domains. **b** Predicted maximum production rate when introduced into a *S. coelicolor* GEM. The x- and y-axis represent the maximal production rate using the metabolic pathway created based on literature or reconstructed with BiGMeC, respectively. **c** This panel shows a comparison of the predicted reaction-knockout targets (x-axis) when using a metabolic pathway created based on literature or with BiGMeC. Similar predictions are shown as green tiles, while incorrect predictions (predicted in either but not both of the two cases) are shown as red tiles. The names of the model reaction IDs are: TKT1: transketolase; ASPT: aspartate ammonia-lyase; FERO: ferroxidase; GLYCL: glycine cleavage system; MCOALY: malyl-CoA lyase; AGT: alanine-glyoxylate aminotransferase; FUM: fumarase; ASPTA: aspartate transaminase; CITMS: (R)-citramalate synthase; ERTHMMOR: 3-isopropylmalate dehydrogenase; CITCla2: (R)-2-Methylmalate hydro-lyase; CITCib: 2-methylmaleate hydratase; GHMT2r: glycine hydroxymethyltransferase; PSERT: phosphoserine transaminase; PGCD: phosphoglycerate dehydrogenase; PSP_L: phosphoserine phosphatase; FDH: formate dehydrogenase

to handle deviations from canonical rules and cases where information is missing (see Materials and Methods). Improvements in determining module function includes identification of bridging modules in trans-AT PKSs and non-extending modules due to the presence of oMT domains [32] (Fig. 1b).

We first assessed the accuracy of the BiGMeC pipeline by comparing its predictions with experimentally characterized and manually curated metabolic pathways. To this end, we compared the substrates, cofactors, and reaction products of each step of the metabolic pathway associated with eight well-characterized BGCs (Fig. 2a, Additional file 1). These BGCs cover a range of BGC classes, including type 1 PKS, trans-AT PKS, NRPS and hybrids, and we believe they provide a test set that is sufficiently diverse to probe the pipeline for its strengths and weaknesses. Overall, BiGMeC appends the

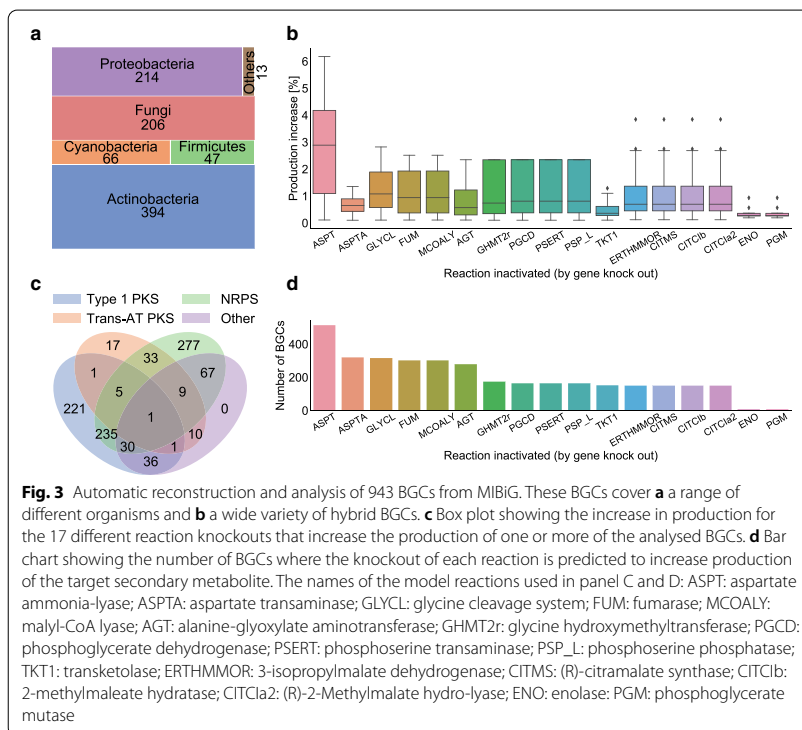
correct metabolic reaction for 72.8% (166/228) of the functional domains in all eight BGCs. Of these functional domains, BiGMeC chooses the correct extender unit for 81.3% (74/91) of the domains extending the peptide or polyketide. For all other domains, including chain initiation, reductive domains, methyltransferases and final tailoring reactions, the accuracy is 67.2% (92/137).

A large number of the incorrect predictions derive from wrong assignments of inactive KR domains by antiSMASH [12]. Across the eight closely inspected BGCs, KR domains are almost always active, but on several occasions antiSMASH predicts that these domains are inactive. The incorrect predictions of KR domain activity are to a large extent associated with adjacent MT domains. Furthermore, this leads to incorrect assignment of the activity of succeeding DH and ER domains because they act on the functional moiety produced by the preceding domain. For the prediction of extender units, most incorrect assignments derive from missing recognition of non-elongating modules caused by inactive KS domains devoid of a conserved histidine residue required for carboxylative condensation [32]. More specifically, only 10 of 16 KS domains are active in the oocydin BGC [32, 41]. Another significant source of incorrect domains is the anabaenopeptin cluster that has two consecutive genes, each having two modules that initiate biosynthesis and perform first chain elongation, respectively, yielding two slightly different variants of the final compound. The BiGMeC pipeline treats these two genes as consecutive steps of the same pathway, and therefore, predicts too many chain elongations in the biosynthesis.

To investigate how much the errors in the constructed metabolic pathways affect model predictions, we introduced both the literature-based and the BiGMeC pathway reconstructions into the consensus GEM of *S. coelicolor* (Sco-GEM) [16] and compared the maximal production rate of the final compound (Fig. 2b). In general, we observe quite similar rates for the eight BGCs (Pearson $\rho = 0.75$, $P = 0.03$), suggesting that the incorrect domains only have a minor impact on the predicted production rates. The offset in the production of leupyrrin likely comes from an incorrect starter unit while the offset in oocydin production is caused by a fairly large error in the predicted number of malonyl-CoA extender units (10 vs. 16).

The anticipated use of the developed pipeline towards strain engineering of expression hosts underscores the need to elucidate if model-based strain designs using BiGMeC-constructed pathways deviate from results using pathways reconstructed according to literature. To this end, we predicted optimal single-reaction knockout mutants that should increase the production rate of the associated product (Fig. 2c). Note that, a reaction knockout is the practical implication of disrupting one or more of the genes encoding the enzyme catalyzing the corresponding reaction. For 6 out of 8 BGCs there is a good overlap between pairwise pathway reconstructions. This includes the cases of tolaasin and geldanamycin, where no knockout target is identified with either of the two pathway reconstructions.

To demonstrate the power of BiGMeC in high-throughput assessment of BGCs, we employed the pipeline on 1883 of the 1923 BGCs in the MIBiG database (version 2.0) [37]. For 40 of the 1923 BGCs, we could not obtain the antiSMASH output file because the link from MIBiG was broken. The 943 (50.1%) metabolic pathways that were successfully reconstructed with BiGMeC cover both fungi and a range of different bacteria



(Fig. 3a). Most clusters are either type 1 PKS, NRPS, or hybrids of these two, and only 77 of the BGCs share similarity with trans-AT PKS (Fig. 3b). The 940 remaining BGCs were not analyzed either because the BGC class was not covered by BiGMeC (such as RiPPs, terpenes, Type 2 and Type 3 PKSs) or because functional modules and domains were lacking in the results from antiSMASH.

We introduced each of the 943 reconstructed pathways into Sco-GEM [16], and predicted single-reaction knockout strategies improving the production of the final pathway product. Surprisingly, only 17 different reactions were suggested as a knockout target in one or more of the 943 *in silico* heterologous expression experiments (Fig. 3c, d). Of these 17 reactions, aspartate transaminase is predicted to provide on average the largest increase in production (Fig. 3c) and is also the most frequently suggested candidate (Fig. 3d). However, the predicted production increase is minor for all of the 17 suggested reactions, including aspartate transaminase, with a maximum increase of 6% relative to the wild-type production rate.

Discussion

To make novel natural product pathways encoded by BGCs accessible to the constraint-based reconstruction and analysis framework, we have developed a pipeline that creates a draft reconstruction of the metabolic pathway encoded by a BGC. This pipeline

outlines the correct metabolic reaction for 72.8% of the functional domains in our test set comprised of 8 experimentally characterized BGC-encoded biosynthetic pathways. One may question whether this accuracy extends to uncharacterized BGCs. In principle, as the pathway reconstruction is solely based on genome mining results from antiSMASH, there should not be a significant difference in accuracy between well-characterized and uncharacterized BGCs. However, as antiSMASH relies on annotation rules learnt from well-characterized BGCs [42, 43], one may anticipate that uncharacterized BGCs that deviate from known canonical rules are less accurately annotated by antiSMASH, and therefore less accurately reconstructed by BiGMeC.

By applying the BiGMeC pipeline to 943 BGCs covering NRPSs, PKSs and NRPS-PKSs hybrids from a wide range of organisms we have demonstrated how the pipeline enables high-throughput assessment of potential candidates for heterologous expression. In an assessment of 943 BGCs, we explored general single-gene knockout strategies for increased heterologous production, and although we identify a set of 17 general targets, none provides a drastic increase in production. This result suggests that multiple knockouts, over-expression of genes, or strategies that perturb regulatory mechanisms are necessary to reroute a large amount of precursors from growth towards secondary metabolism, at least in the organism *S. coelicolor*.

Although the accuracy of the BiGMeC pipeline is sufficient to make biologically relevant pathway reconstructions, this work has also revealed aspects where there is room for further improvement. Incorrect assignment of KS and KR domains as active or inactive is a large source of error in PKS metabolic pathways, and incorporation of the recently developed transATor algorithm would provide an improvement in this context [44]. Synthesis of rare precursors and tailoring of the polyketide or peptide succeeding the release from the multidomain enzyme complex are two other features with opportunity for improvement. Although the genes encoding enzymes responsible for the synthesis of rare precursors or for the post-release tailoring steps usually are contained in the BGC, neither their exact function nor their functional order can be accurately predicted. Therefore, the current pipeline relies in certain aspects on assumptions and heuristics that apply in general, but with several exceptions. However, with a continuous improvement in algorithms for annotation and identification of BGCs [12, 44, 45] and increased experimental characterization [37], current generalisations can develop into more accurate pathway reconstructions that encompass a larger range of deviations from canonical rules. Furthermore, as the knowledgebase and algorithms for annotation of iterative PKSs and ribosomally synthesised and post-translationally modified peptides improves [46, 47], these types of BGCs represent obvious targets for further development. Other possible targets include terpenes, alkaloids and glycosides, frequently encoded in plant and fungal genomes [48–50], or polysaccharides which are of large value in dairy industry [51] and medical applications [52], and the most abundant class of prokaryotic BGCs [5]. Nevertheless, accurate pathway reconstruction for these classes of BGCs will require accurate descriptions of the biosynthetic rules encoded in the gene clusters. In this context, tailoring reactions and post-translational modifications represent particular challenges. Further improvement should also aim to accept the output from other annotation software, such as PRISM [53].

Conclusion

The BiGMeC pipeline is, to our knowledge, the first tool for automatic metabolic pathway reconstruction specifically targeting PKS and NRPS BGCs. Although the reconstructed pathways are not able to capture the entire diversity seen in the biosynthesis of NRPSs and PKSs [30, 32], the predicted production rates and reaction knockout targets are comparable to predictions provided using manually reconstructed pathways. Furthermore, the pipeline can aid model reconstruction efforts, both as a decent starting point for further manual curation and as a complement to standard model-reconstruction pipelines [54]. This is in particular relevant for organisms with a rich secondary metabolism, such as the *Actinobacteria* which are of utmost interest in drug discovery. We anticipate that the pipeline presented here can increase the use of GEMs in this context, e.g. to screen different combinations of BGCs and expression hosts or, as shown in this work, to explore strain-engineering opportunities. The pipeline is developed in an open source environment on GitHub and we encourage interested readers to engage in future development through pull request or by raising issues. We also encourage developers of genome mining tools and databases to converge towards standardized and consistent file formats, such as the Minimum Information about a Biosynthetic Gene Cluster (MIBiG) initiative [37]. This will ease the development and maintenance of downstream pipelines such as BiGMeC, and promote integration of data from different genome mining tools. This is intended as a reminder rather than a criticism of existing software.

Materials and methods

Software implementation

We developed BiGMeC to translate information about PKS and NRPS BGCs to detailed outlines of the metabolic reactions governing the production of the associated secondary metabolites. The BiGMeC software and all other associated scripts are implemented in Python 3 and publicly available at <https://github.com/AlmaasLab/BiGMeC>. BiGMeC runs from a command-line interface and takes an annotated NRPS or PKS BGC in the format of a region-specific GenBank file as produced by antiSMASH 5.1 [12]. It leverages the included gene, domain, and module information to make a description of the enzymatic reactions encoded by the BGC, including substrate and co-factor usage (Fig. 1a). BiGMeC uses a reference model as a library of metabolites and reactions, and in the current work, we have used Sco-GEM version 1.2.1, the consensus *S. coelicolor* GEM [16]. This model was obtained from <https://github.com/SysBioChalmers/Sco-GEM>.

The BiGMeC pipeline first parses information about the location and annotation of the genes and modules as annotated by antiSMASH from the GenBank file (Fig. 1). If available, the gene information includes strand, secondary metabolism Clusters of Orthologous Groups (smCOG) annotation [55], type of gene, extender unit, annotated functional domains and if the gene is a core gene or not. The core genes synthesize the core structure of the PKS or NRPS molecule. The module information contains details about the type of module and its functional domains. Then, the pipeline assesses the presence and order of domains not included in a module, e.g. special load or bridging modules (in trans-AT PKS, Fig. 1b) [32], and combines these domains into functional modules when possible. The peptide or polyketide backbone is subsequently constructed

Table 1 List of domains and associated reactions as implemented in BiGMeC

Abbrev.	Name	Note	Reaction
A	Adenylation	Activates and attaches AA to PCP	$ATP + AA + PCP \rightarrow AA-PCP + AMP + P_{ii}$
ACP	Acyl carrier protein	Facilitates transport in PKSs	
AT	Acyltransferase	Loads extender unit onto ACP	$Acyl-CoA + ACP \rightarrow Acyl-ACP + CoA$
C	Condensation	Elongates the peptide by condensation	$AA-PCP + X_n \rightarrow X_{n+1} + H_2O + ACP$
CAL	Coenzyme A ligase	Catalyzes the incorporation of different starter units, e.g. fatty acids, AHBA, and shikimic acid [32, 56, 57]	$Y + PCP \rightarrow Y-PCP + H_2O$
cMT	Carbon methyltransferase	Methylates peptide/polyketide	$SAM + X_n \rightarrow X_n + SAH$
Cy	Heterocyclization	Elongates the peptide by condensation and cyclization	$AA-PCP + X_n \rightarrow X_{n+1} + H_2O + ACP$
DH	Dehydratase	Forms double bond by removal of H_2O	$X_n \rightarrow X_n + H_2O$
E	Epimerase	Stereochemical inversion	$X_n \rightarrow X_n$
ECH	Enoyl-CoA hydratase/isomerase	Not able to discriminate, so BiGMeC assumes isomerase	$X_n \rightarrow X_n$
ER	Enoyl reductase	Reduces double bond formed by the DH domain to a methylene group	$NADPH + H^+ + X_n \rightarrow X_n + NADP^+$
FkbH	FkbH-like domain	Domain in an alternative loading module. Dephosphorylates 1,3-bpg [58]	$1,3-bpg + ACP \rightarrow D-lactate-ACP + 2 P_i$
GNAT	GCN5-related N-acetyl transferase	Alt. load module that decarboxylates malonyl-CoA and adds acetyl group to ACP [59]	$Malonyl-CoA + ACP \rightarrow Acetyl-ACP + CoA + CO_2$
KR	Keto reductase	Reduces carbonyl group to hydroxyl group	$NADPH + H^+ + X_n \rightarrow X_n + NADP^+$
KS	Keto synthase	Appends extender unit to polyketide	$Acyl-ACP + X_n \rightarrow X_{n+1} + CO_2 + ACP$
nMT	Nitrogen methyltransferase	Methylates peptide/polyketide	$SAM + X_n \rightarrow X_n + SAH$
oMT	Oxygen methyltransferase	Methylates peptide/polyketide	$SAM + X_n \rightarrow X_n + SAH$
PCP	Peptidyl carrier domain	Facilitates transport in NRPSs	
TD	Thioester reductase	Releases product from ACP/PCP	$NADPH + H^+ + X_n \rightarrow \text{detached product} + NADP^+$
TE	Thioesterase	Releases product from ACP/PCP	$H_2O + X_n \rightarrow \text{detached product}$

The peptide or polyketide backbone is referred to as X_n , and in reactions that extend the backbone we refer to the elongated backbone as X_{n+1}

AA, generic amino acid; 1,3-bpg, 1,3 biphosphoglycerate; CoA, Coenzyme A; P_{ii} , diphosphate; SAH, S-Adenosyl-L-homocysteine; SAM, S-Adenosyl methionine; Y, generic starter unit

based on the order of the identified domains and the function of each domain within each module. Although NRPS and type 1 PKS modules can be iterative, we here assume that the selected BGCs are modular such that each module only performs one chain elongation. The reactions associated with the functional domains are listed in Table 1. Domains in the BGC that are not contained in a module are assumed to not affect the backbone structure. If a terminating domain (thioesterase or thioester reductase) domain is encountered, no further chain elongations are carried out. The activity of reducing domains (DH, ER, KR) are based on the annotation of the KR domain from antiSMASH. Tailoring reactions post PKS synthesis are predicted from the smCOG annotations of each gene. The currently implemented tailoring reactions relate to the

smCOGs 1256, 1084, 1002, 1109 and 1062 and includes glycosylation, glycosyltransferase and incorporation of 2-Amino-3-hydroxycyclopent-2-enone (Additional file 2).

Rare extender units appear in both PKS and NRPS biosynthesis. The synthesis of rare extender units is usually carried out by genes in the BGC [60], and we therefore include the synthesis of the most common rare extender units (not in the reference library) when necessary. This includes hydroxyphenylglycine, beta-hydroxytyrosine, 2-aminobutyric acid, pipercolic acid, dihydroxyphenylglycine and 3-amino-5-hydroxybenzoate [56]. Synthesis of the rare extender unit methoxymalonyl-ACP [60] is based on the presence of genes with specific smCOG annotations (Additional file 2). For the remaining rare extender units, or in the case of missing information or nonspecific antiSMASH annotation, we use a conservative approach where a generic amino acid is used as the extender unit in NRPS modules and malonyl-CoA is used in PKS modules. In the case of using a generic amino acid as the extender unit, we add a set of pseudo-reactions that can convert every proteogenic amino acid into this generic molecule to ensure that the biosynthetic pathway is functional.

The pipeline also handles a number of deviations from the canonical rules, for example the deactivation of the KS domain often seen in modules containing O-methyltransferases [32]. Furthermore, it is found that the presence of a C domain in the initiating NRPS module acylates the initial amino acid [31, 61]. Both in tolaasin [62] and surfactin, currently the best studied example of this type of NRPS initiation, the acylating agent is a CoA-activated β -hydroxy fatty acid [61, 63]. It is likely that the C-domain has a strong selectivity for a specific acylating agent, but since this specificity is not identified by antiSMASH we use a generic fatty acid molecule. A third example of exceptions that are handled by BiGMeC is bridging modules in trans-AT PKSs where the KS domain is encoded in the first gene and the DH and ACP domains follow immediately on the second gene. These modules are called dehydratase docking domains (DHD) and are usually not active [32].

Evaluation of the BiGMeC pipeline

To evaluate how well biosynthetic pathways can be constructed solely based on antiSMASH data we compared BiGMeC-constructed pathways with literature-based reconstructions for 8 different BGCs, covering different species and classes of BGCs (Additional file 1). The 8 BGCs were (MIBiG ID in parenthesis): bafilomycin from *Streptomyces lohi* [64–66] (BGC0000028), geldanamycin from *Streptomyces hygroscopicus* [67–69] (BGC0000066), diffidin from *Bacillus velezensis* FZB42 [70, 71] (BGC0000176), oocycin from *Serratia plymuthica* [32, 41] (BGC0001032), oxazolomycin from *Streptomyces albus* [71, 72] (BGC0001106), leupyrrin from *Sorangium cellulosum* [73] (BGC0000380), anabaenopeptin from *Anabaena* sp. 90 [74] (BGC0000302) and tolaasin from *Pseudomonas costantinii* [62] (BGC0000447). For each domain in each of the 8 different BGCs we compared the BiGMeC-constructed reaction with the *real* reaction, i.e. the associated reaction as described in the literature. When clearly defined in the literature, tailoring reactions were included, but we focused on the synthesis of the core peptide/polyketide. The very complex tailoring of leupyrrin [73] was not included.

An initial evaluation was performed by counting the number of correct domains (Fig. 2a). The total number of domains include all domains either predicted by BiGMeC or described in the literature, and the correct predictions include both true positives and true negatives. Next, we incorporated the BiGMeC and literature-based

pathway reconstructions into Sco-GEM and predicted the maximum production rate of the secondary metabolite produced by each pathway (Fig. 2b). To do so, we performed Flux Balance Analysis (FBA) [75, 76] in cobrapy [39] with the final reaction of the BGC encoded pathway as objective and with growth limited to minimum 90% of the maximum value. The growth and production were simulated in a growth medium with glucose and ammonium as the sole carbon and nitrogen sources, respectively, and with a maximum glucose uptake rate of $0.8 \text{ mmol gDW}^{-1} \text{ h}^{-1}$. We did not constrain the uptake of ammonium, sulphate, phosphate, oxygen and metal ions. Finally, using both the BiGMeC and literature-based pathway reconstructions, we predicted reaction inactivation targets (by gene knockout) that would increase the production of the associated compound, with a maximum growth rate reduction of 50% (Fig. 2c). We limited the set of possible reaction targets to non-essential gene-annotated reactions. The search for optimal knockouts was carried out in a brute-force manner: we conducted an iterative knockout of each reaction (within the predefined set of possible reactions) and, first used FBA to predict the maximum growth of the mutant phenotype, and secondly predict the maximum production rate at 99.9% of the knockout-mutant's maximum growth rate. All knockouts that resulted in more than 0.1% increase in production rate compared to the wild-type were considered knockout candidates.

Large-scale reconstruction of BGC pathways

To demonstrate the value and efficiency enabled by BiGMeC we applied this pipeline to all relevant BGCs from the MIBiG database [37]. To get the antiSMASH-generated output for all BGCs in MIBiG we automatically downloaded all GenBank-files with a url on the form: <https://mibig.secondarymetabolites.org/repository/BGC0000001/generated/BGC0000001.1.region001.gbk>, with the MIBiG ID ranging from BGC0000001 to BGC0002057. The MIBiG database currently reports on a total of 1923 BGCs but due to different reasons (e.g. missing entries) we could only obtain the antiSMASH result for 1883 of the entries. For all BGCs at least annotated to either type 1 PKS, trans-AT PKS or NRPS we used the BiGMeC pipeline to reconstruct the corresponding metabolic pathway. We predicted optimal knockout strategies for each of successfully constructed pathway using the same procedure as described for the 8 BGCs used to evaluate the BiGMeC pipeline.

Supplementary Information

The online version contains supplementary material available at <https://doi.org/10.1186/s12859-021-03985-0>.

Additional file 1. Details on tailoring reactions and synthesis of the rare extender unit methoxymalonyl-ACP, as well as a description of the analysis used to develop the heuristics that indicate the presence of these reactions from smCOG annotations.

Additional file 2. Detailed comparison of 8 BGCs for evaluation the accuracy of the BiGMeC pipeline.

Acknowledgements

Not applicable.

Authors' contributions

Conceptualization, SS, EA; Methodology and Software, SS, FF; Validation and Formal Analysis, SS, FF; Data curation SS, FF; Writing: Original Draft, SS, FF; Reviewing and editing, SS, EA, FF, AW; Visualization SS; Supervision SS, EA; Project Administration, AW, EA; Funding Acquisition, AW, EA. All authors read and approved the final manuscript.

Funding

This research was conducted within the project INBioPharm of the Center for Digital Life Norway (Research Council of Norway grant #248885), with additional support of SINTEF internal funding.

Availability of data and materials

The BiGMeC pipeline and the data analysed/generated during the current study is available at <https://github.com/AlmaasLab/BiGMeC>. We have also deposited the latest version of the repository to Zenodo (<https://doi.org/10.5281/zenodo.4434667>) to ensure persistent access.

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Competing interests

The authors declare that they have no competing interests.

Author details

¹ Department of Biotechnology and Food Science, NTNU - Norwegian University of Science and Technology, Sem Sælands vei 8, 7034 Trondheim, Norway. ² Department of Biotechnology and Nanomedicine, SINTEF Industry, Richard Birkelands vei 3, 7034 Trondheim, Norway. ³ K.G. Jebsen Center for Genetic Epidemiology, NTNU - Norwegian University of Science and Technology, Håkon Jarls gate 11, 7030 Trondheim, Norway.

Received: 26 November 2020 Accepted: 18 January 2021

Published online: 23 February 2021

References

- Clardy J, Fischbach MA, Walsh CT. New antibiotics from bacterial natural products. *Nat Biotechnol.* 2006;24(12):1541–50.
- Demain AL, Sanchez S. Microbial drug discovery: 80 years of progress. *J Antibiot.* 2009;62(1):5–16.
- Cantrell CL, Dayan FE, Duke SO. Natural products as sources for new pesticides. *J Nat Prod.* 2012;75(6):1231–42.
- Rokas A, Mead ME, Steenwyk JL, Raja HA, Oberlies NH. Biosynthetic gene clusters and the evolution of fungal chemodiversity. *Nat Prod Rep.* 2020;37:868–78.
- Cimermancic P, Medema MH, Claesen J, Kurita K, Brown LCW, Mavrommatis K, Pati A, Godfrey PA, Koehrsen M, Clardy J, et al. Insights into secondary metabolism from a global analysis of prokaryotic biosynthetic gene clusters. *Cell.* 2014;158(2):412–21.
- Bentley SD, Chater KF, Cerdeño-Tárraga A-M, Challis GL, Thomson N, James KD, Harris DE, Quail MA, Kieser H, Harper D, et al. Complete genome sequence of the model actinomycete *Streptomyces coelicolor* a3 (2). *Nature.* 2002;417(6885):141–7.
- Ikeda H, Ishikawa J, Hanamoto A, Shinose M, Kikuchi H, Shiba T, Sakaki Y, Hattori M, Ōmura S. Complete genome sequence and comparative analysis of the industrial microorganism *Streptomyces avermitilis*. *Nat Biotechnol.* 2003;21(5):526–31.
- Harvey AL, Edrada-Ebel R, Quinn RJ. The re-emergence of natural products for drug discovery in the genomics era. *Nat Rev Drug Discov.* 2015;14(2):111–29.
- Xu M, Wright GD. Heterologous expression-facilitated natural products' discovery in actinomycetes. *J Ind Microbiol Biotechnol.* 2019;46(3–4):415–31.
- Myronovskiy M, Luzhetskyy A. Heterologous production of small molecules in the optimized *Streptomyces* hosts. *Nat Prod Rep.* 2019;36(9):1281–94.
- Kim HU, Blin K, Lee SY, Weber T. Recent development of computational resources for new antibiotics discovery. *Curr Opin Microbiol.* 2017;39:113–20.
- Blin K, Shaw S, Steinke K, Villebro R, Ziemert N, Lee SY, Medema MH, Weber T. Antismash 5.0: updates to the secondary metabolite genome mining pipeline. *Nucleic Acids Res.* 2019;47(W1):81–7.
- Huo L, Hug JJ, Fu C, Bian X, Zhang Y, Müller R. Heterologous expression of bacterial natural product biosynthetic pathways. *Nat Prod Rep.* 2019;36(10):1412–36.
- Sekurova ON, Schneider O, Zotchev SB. Novel bioactive natural products from bacteria via bioprospecting, genome mining and metabolic engineering. *Microb biotechnol.* 2019;12(5):828–44.
- Nah H-J, Pyeon H-R, Kang S-H, Choi S-S, Kim E-S. Cloning and heterologous expression of a large-sized natural product biosynthetic gene cluster in *Streptomyces* species. *Front Microbiol.* 2017;8:394.
- Sulheim S, Kumelji T, van Dissel D, Salehzadeh-Yazdi A, Du C, van Wezel GP, Nieselt K, Almaas E, Wentzel A, Kerkhoven EJ. Enzyme-constrained models and omics analysis of *Streptomyces coelicolor* reveal metabolic changes that enhance heterologous production. *iScience.* 2020;23(9):101525.
- Ke J, Yoshikuni Y. Multi-chassis engineering for heterologous production of microbial natural products. *Curr Opin Biotechnol.* 2020;62:88–97.

18. Famili I, Förster J, Nielsen J, Palsson BO. *Saccharomyces cerevisiae* phenotypes can be predicted by using constraint-based analysis of a genome-scale reconstructed metabolic network. *Proc Natl Acad Sci*. 2003;100(23):13134–9.
19. Gu C, Kim GB, Kim WJ, Kim HU, Lee SY. Current status and applications of genome-scale metabolic models. *Genome Biol*. 2019;20(1):121.
20. Mohite OS, Weber T, Kim HU, Lee SY. Genome-scale metabolic reconstruction of actinomycetes for antibiotics production. *Biotechnol J*. 2019;14(11):1800377.
21. Barka EA, Vatsa P, Sanchez L, Gaveau-Vaillant N, Jacquard C, Klenk H-P, Clément C, Ouhdouch Y, van Wezel GP. Taxonomy, physiology, and natural products of actinobacteria. *Microbiol Mol Biol Rev*. 2016;80(1):1–43.
22. Wang H, Marcišauskas S, Sánchez BJ, Domenzain I, Hermansson D, Agren R, Nielsen J, Kerkhoven EJ. Raven 2.0: a versatile toolbox for metabolic network reconstruction and a case study on *Streptomyces coelicolor*. *PLoS Comput Biol*. 2018;14(10):1006541.
23. Borodina I, Siebring J, Zhang J, Smith CP, van Keulen G, Dijkhuizen L, Nielsen J. Antibiotic overproduction in *Streptomyces coelicolor* a3 (2) mediated by phosphofructokinase deletion. *J Biol Chem*. 2008;283(37):25186–99.
24. Huang D, Li S, Xia M, Wen J, Jia X. Genome-scale metabolic network guided engineering of *Streptomyces tsukubaensis* for fks506 production improvement. *Microb Cell Factories*. 2013;12(1):1–18.
25. Kumelj T, Sulheim S, Wentzel A, Almaas E. Predicting strain engineering strategies using iks1317: a genome-scale metabolic model of *Streptomyces coelicolor*. *Biotechnol J*. 2019;14(4):1800180.
26. Doroghazi JR, Metcalf WW. Comparative genomics of actinomycetes with a focus on natural product biosynthetic genes. *BMC Genom*. 2013;14(1):611.
27. Masschelein J, Jenner M, Challis GL. Antibiotics from gram-negative bacteria: a comprehensive overview and selected biosynthetic highlights. *Nat Prod Rep*. 2017;34(7):712–83.
28. Bozhüyük KA, Micklefield J, Wilkinson B. Engineering enzymatic assembly lines to produce new antibiotics. *Curr Opin Microbiol*. 2019;51:88–96.
29. Cane DE, Walsh CT, Khosla C. Harnessing the biosynthetic code: combinations, permutations, and mutations. *Science*. 1998;282(5386):63–8.
30. Challis GL, Naismith JH. Structural aspects of non-ribosomal peptide biosynthesis. *Curr Opin Struct Biol*. 2004;14(6):748–56.
31. Fischbach MA, Walsh CT. Assembly-line enzymology for polyketide and nonribosomal peptide antibiotics: logic, machinery, and mechanisms. *Chem Rev*. 2006;106(8):3468–96.
32. Helfrich EJ, Piel J. Biosynthesis of polyketides by trans-acting polyketide synthases. *Nat Prod Rep*. 2016;33(2):231–316.
33. Keatinge-Clay AT. The structures of type I polyketide synthases. *Nat Prod Rep*. 2012;29(10):1050–73.
34. Mootz HD, Schwarzer D, Marahiel MA. Ways of assembling complex natural products on modular nonribosomal peptide synthetases. *ChemBioChem*. 2002;3(6):490–504.
35. Fisch KM. Biosynthesis of natural products by microbial iterative hybrid pks-nrps. *RSC Adv*. 2013;3(40):18228–47.
36. Herbst DA, Townsend CA, Maier T. The architectures of iterative type I pks and fas. *Nat Prod Rep*. 2018;35(10):1046–69.
37. Kautsar SA, Blin K, Shaw S, Navarro-Muñoz JC, Terlouw BR, van der Hooft JJ, Van Santen JA, Tracanna V, Suarez Duran HG, Pascal Andreu V, et al. Mibig 2.0: a repository for biosynthetic gene clusters of known function. *Nucleic Acids Res*. 2020;48(D1):454–8.
38. Zhang JJ, Tang X, Moore BS. Genetic platforms for heterologous expression of microbial natural products. *Nat Prod Rep*. 2019;36(9):1313–32 (Publisher: The Royal Society of Chemistry).
39. Ebrahim A, Lerman JA, Palsson BO, Hyduke DR. Cobrapy: constraints-based reconstruction and analysis for python. *BMC Syst Biol*. 2013;7(1):74.
40. Heirendt L, Arreckx S, Pfau T, Mendoza SN, Richelle A, Heinken A, Haraldsdóttir HS, Wachowiak J, Keating SM, Vlasov V, et al. Creation and analysis of biochemical constraint-based models using the cobra toolbox v. 3.0. *Nat Protoc*. 2019;14(3):639–702.
41. Matilla MA, Stöckmann H, Leeper FJ, Salmond GP. Bacterial biosynthetic gene clusters encoding the anti-cancer haterumalide class of molecules biogenesis of the broad spectrum antifungal and anti-oomycete compound, oocycin A. *J Biol Chem*. 2012;287(46):39125–38.
42. Blin K, Kim HU, Medema MH, Weber T. Recent development of antimash and other computational approaches to mine secondary metabolite biosynthetic gene clusters. *Brief Bioinform*. 2019;20(4):1103–13.
43. Medema MH, Fischbach MA. Computational approaches to natural product discovery. *Nat Chem Biol*. 2015;11(9):639.
44. Helfrich EJ, Ueoka R, Dolev A, Rust M, Meoded RA, Bhushan A, Califano G, Costa R, Gugger M, Steinbeck C, et al. Automated structure prediction of trans-acyltransferase polyketide synthase products. *Nat Chem Biol*. 2019;15(8):813–21.
45. Kjærboelling I, Mortensen UH, Vesth T, Andersen MR. Strategies to establish the link between biosynthetic gene clusters and secondary metabolites. *Fungal Genet Biol*. 2019;130:107–21.
46. Wang B, Guo F, Huang C, Zhao H. Unraveling the iterative type I polyketide synthases hidden in streptomyces. *Proc Natl Acad Sci*. 2020;117(15):8449–54.
47. Kloosterman AM, Cimermanic P, Elsayed SS, Du C, Hadjithomas M, Donia MS, Fischbach MA, van Wezel GP, Medema MH. Expansion of RIPP biosynthetic space through integration of pan-genomics and machine learning uncovers a novel class of lantibiotics. *PLoS Biol*. 2020;18(12):3001026.
48. Kautsar SA, Suarez Duran HG, Blin K, Osbourn A, Medema MH. Plantismash: automated identification, annotation and expression analysis of plant biosynthetic gene clusters. *Nucleic Acids Res*. 2017;45(W1):55–63.
49. Li YF, Tsai KJ, Harvey CJ, Li JJ, Ary BE, Berlew EE, Boehman BL, Findley DM, Friant AG, Gardner CA, et al. Comprehensive curation and analysis of fungal biosynthetic gene clusters of published natural products. *Fungal Genet Biol*. 2016;89:18–28.
50. Nützmann H-W, Huang A, Osbourn A. Plant metabolic clusters—from genetics to genomics. *New Phytologist*. 2016;211(3):771–89.
51. Duboc P, Mollet B. Applications of exopolysaccharides in the dairy industry. *Int Dairy J*. 2001;11(9):759–68.

52. Moscovici M. Present and future medical applications of microbial exopolysaccharides. *Front Microbiol.* 2015;6:1012.
53. Skinnider MA, Merwin NJ, Johnston CW, Magarvey NA. Prism 3: expanded prediction of natural product chemical structures from microbial genomes. *Nucleic Acids Res.* 2017;45(W1):49–54.
54. Mendoza SN, Olivier BG, Molenaar D, Teusink B. A systematic assessment of current genome-scale metabolic reconstruction tools. *Genome Biol.* 2019;20(1):1–20.
55. Medema MH, Blin K, Cimermancic P, de Jager V, Zakrzewski P, Fischbach MA, Weber T, Takano E, Breitling R. Antismash: rapid identification, annotation and analysis of secondary metabolite biosynthesis gene clusters in bacterial and fungal genome sequences. *Nucleic Acids Res.* 2011;39(suppl-2):339–46.
56. Floss HG, Yu T-W, Arakawa K. The biosynthesis of 3-amino-5-hydroxybenzoic acid (ahba), the precursor of mc 7 n units in ansamycin and mitomycin antibiotics: a review. *J Antibiot.* 2011;64(1):35–44.
57. Fritzlner JM, Zhu G. Functional characterization of the acyl-[acyl carrier protein] ligase in the cryptosporidium parvum giant polyketide synthase. *Int J Parasitol.* 2007;37(3–4):307–16.
58. Zhang F, He H-Y, Tang M-C, Tang Y-M, Zhou Q, Tang G-L. Cloning and elucidation of the fr901464 gene cluster revealing a complex acyltransferase-less polyketide synthase using glycerate as starter units. *J Am Chem Soc.* 2011;133(8):2452–62.
59. Gu L, Geders TW, Wang B, Gerwick WH, Håkansson K, Smith JL, Sherman DH. Gnat-like strategy for polyketide chain initiation. *Science.* 2007;318(5852):970–4.
60. Chan YA, Podevels AM, Kevany BM, Thomas MG. Biosynthesis of polyketide synthase extender units. *Nat Prod Rep.* 2009;26(1):90–114.
61. Kraas FI, Helmetag V, Wittmann M, Strieker M, Marahiel MA. Functional dissection of surfactin synthetase initiation module reveals insights into the mechanism of lipoinitiation. *Chem Biol.* 2010;17(8):872–80.
62. Scherlach K, Lackner G, Graupner K, Pidot S, Bretschneider T, Hertweck C. Biosynthesis and mass spectrometric imaging of tolaasin, the virulence factor of brown blotch mushroom disease. *ChemBioChem.* 2013;14(18):2439–43.
63. Steller S, Sokoll A, Wilde C, Bernhard F, Franke P, Vater J. Initiation of surfactin biosynthesis and the role of the srfD-thioesterase protein. *Biochemistry.* 2004;43(35):11331–43.
64. Zhang W, Fortman JL, Carlson JC, Yan J, Liu Y, Bai F, Guan W, Jia J, Matainaho T, Sherman DH, et al. Characterization of the baflomycin biosynthetic gene cluster from streptomyces lohii. *Chembiochem Eur J Chem Biol.* 2013;14(3):301.
65. Nara A, Hashimoto T, Komatsu M, Nishiyama M, Kuzuyama T, Ikeda H. Characterization of baflomycin biosynthesis in kitasatospora setae km-6054 and comparative analysis of gene clusters in actinomycetales microorganisms. *J Antibiot.* 2017;70(5):616–24.
66. Li Z, Du L, Zhang W, Zhang X, Jiang Y, Liu K, Men P, Xu H, Fortman JL, Sherman DH, et al. Complete elucidation of the late steps of baflomycin biosynthesis in streptomyces lohii. *J Biol Chem.* 2017;292(17):7095–104.
67. Patel K, Piagentini M, Rascher A, Tian Z-Q, Buchanan GO, Regentin R, Hu Z, Hutchinson C, McDaniel R. Engineered biosynthesis of geldanamycin analogs for hsp90 inhibition. *Chem Biol.* 2004;11(12):1625–33.
68. Rascher A, Hu Z, Viswanathan N, Schirmer A, Reid R, Nierman WC, Lewis M, Hutchinson CR. Cloning and characterization of a gene cluster for geldanamycin production in streptomyces hygroscopicus nrrl 3602. *FEMS Microbiol Lett.* 2003;218(2):223–30.
69. Rascher A, Hu Z, Buchanan GO, Reid R, Hutchinson CR. Insights into the biosynthesis of the benzoquinone ansamycins geldanamycin and herbimycin, obtained by gene sequencing and disruption. *Appl Environ Microbiol.* 2005;71(8):4862–71.
70. Chen X-H, Vater J, Piel J, Franke P, Scholz R, Schneider K, Koumoutsis A, Hitzeroth G, Grammel N, Strittmatter AW, et al. Structural and functional characterization of three polyketide synthase gene clusters in bacillus amyloliquefaciens fzb 42. *J Bacteriol.* 2006;188(11):4024–36.
71. Piel J. Biosynthesis of polyketides by trans-acting polyketide synthases. *Nat Prod Rep.* 2010;27(7):996–1047.
72. Zhao C, Ju J, Christenson SD, Smith WC, Song D, Zhou X, Shen B, Deng Z. Utilization of the methoxymalonyl-acyl carrier protein biosynthesis locus for cloning the oxazolomycin biosynthetic gene cluster from streptomyces albus ja3453. *J Bacteriol.* 2006;188(11):4142–7.
73. Kopp M, Irschik H, Gemperlein K, Buntin K, Meiser P, Weissman KJ, Bode HB, Müller R. Insights into the complex biosynthesis of the leupyrrins in sorangium cellululosum so ce690. *Mol Biosyst.* 2011;7(5):1549–63.
74. Rouhiainen L, Jokela J, Fewer DP, Urmann M, Sivonen K. Two alternative starter modules for the non-ribosomal biosynthesis of specific anabaenopeptin variants in anabaena (cyanobacteria). *Chem Biol.* 2010;17(3):265–73.
75. Fell DA, Small JR. Fat synthesis in adipose tissue. an examination of stoichiometric constraints. *Biochem J.* 1986;238(3):781–6.
76. Orth JD, Thiele I, Palsson BØ. What is flux balance analysis? *Nat Biotechnol.* 2010;28(3):245–8.

Publisher's Note

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

Supplemental Material 1

Automatic reconstruction of metabolic pathways from identified biosynthetic gene clusters

Snorre Sulheim, Fredrik A. Fossheim, Alexander Wentzel and Eivind Almaas

Tailoring reactions

We surveyed gene clusters in MIBiG [1] to explore how specific smCOG gene annotations could be used to determine the tailoring reactions associated with the biosynthesis of a polyketide or non-ribosomal peptide synthetase: we compared the presence of specific secondary metabolism Clusters of Orthologous Groups (smCOGs) [2] annotations with the presence of different tailoring reactions as described in the literature. This led to three fairly robust heuristic that were implemented into the BiGMeC pipeline (Table S1). The data that these heuristics are based on are further described in separate sections below.

Table S1: Boolean logic used to determine tailoring reactions from smCOG gene annotation. The peptide or polyketide that is modified by the tailoring reactions is denoted by X_n , both prior and subsequent to the tailoring. Other abbreviations used in the table: bpg: 1,3-biphopshoglycerate; P_i : phosphate; P_{ii} : diphosphate; CoA: Coenzyme A.

smCOG logic	Reaction
1256 and 1084	$bpg + NADH + H^+ + X_n \rightarrow X_n + 2 P_i + NAD^+$
1002 and 1109	$glycine + succinyl-CoA + ATP + X_n \rightarrow X_n + AMP + P_{ii} + CO_2 + H_2O + CoA$
1062	$glucose\ 6-phosphate + X_n \rightarrow X_n + P_i + H^+$

Addition of glycerate

Tailoring by addition of glycerate from 1,3-biphopshoglycerate was identified from the presence of the smCOGs 1256 (FkbH like protein) and 1084 (3-oxoacyl-(acyl carrier protein) synthase III). Across all BGCs in MIBiG, this heuristic provides a correct tailoring reaction in 10 out of 11 cases (Table S2). The single error derives from BGC0000082 which still incorporates a glycerate unit, but lacking the smCOG 1084 annotation.

Incorporation of 2-Amino-3-hydroxycyclopent-2-enone

The tailoring reaction that incorporate 2-Amino-3-hydroxycyclopent-2-enone synthesized from succinyl-CoA and glycine [12], was identified from the

Table S2: Listing of BGCs in MIBiG used to determine when the addition of glycerate is used to tailor the polyketide or peptide. These are all BGCs in MIBiG containing the FkbH domain and adds glycerate. Of these 11 BGCs, 10 are annotated with the smCOG 1084.

MIBiG ID	Product	1084 present	Reference
BGC0000001	Abyssomycin	Yes	[3]
BGC0000036	Chlorothricin	Yes	[4]
BGC0000082	Kijanamicin	No	[5]
BGC0000133	Quartromicin	Yes	[6]
BGC0000140	4H3H2HMe2HF5	Yes	[6]
BGC0000162	Tetrocarcin	Yes	[7]
BGC0000164	Tetronomycin	Yes	[8]
BGC0001004	Lobophorin B	Yes	[9]
BGC0001183	Lobophorin A	Yes	[9]
BGC0001204	Versipelostatin	Yes	[10]
BGC0001288	Maklamicin	Yes	[11]

presence of the smCOGs of 1109 (8-amino-7-oxononanoate synthase) and a neighboring 1002 (AMP-dependent ligase and synthetase). This pair of smCOG annotations were present in 8 BGCs in MIBiG, of which 5 features a tailoring reaction that incorporates 2-Amino-3-hydroxycyclopent-2-enone (Table S3). For one of the BGCs (BGC0001420) the available literature was not sufficient to make a clear decision about the compound tailoring. The remaining two BGCs (BGC0000091 and BGC0001063) feature a similar tailoring reaction where the succinyl-CoA precursor is replaced by several malonyl-CoA units [13, 14].

Glycosylation

Glycosylation of the peptide or polyketide is identified by the presence of smCOG 1062 (glycosyltransferase), and as one might expect the number of incorporated sugar monomers increase with increasing number of smCOG 1062 annotations (Table S4). Based on 40 BGC synthesis pathways that incorporate sugar monomers (we consider Komodoquinone B as an outlier), we find a significant correlation between the number of incorporated sugar monomers and the number of smCOG 1062 Pearson $\rho = 0.78$, $P = 3e - 9$. Based on this result the BiGMeC pipeline therefore assumes that all glycosyltransferases are active, i.e. a 1 to 1 relationship between the number of smCOG 1062 and active glycosyltransferase tailoring reactions.

Table S3: Listing of BGCs in MIBiG used to determine when the peptide or polyketide is tailored by the incorporation of 2-Amino-3-hydroxycyclopent-2-enone synthesized from succinyl-CoA and glycine [12]. All of these BGCs are annotated with the smCOG 1002 and 1109 on adjacent genes. The "Correct" column indicate if these synthesis of the molecules associated with each BGC incorporates 2-Amino-3-hydroxycyclopent-2-enone.

MIBiG ID	Product	Correct	Reference
BGC0000028	Bafilomycin	Yes	[15]
BGC0000091	Marineosin	No	[13]
BGC0000187	Asukamycin	Yes	[16]
BGC0000213	Colabomycin	Yes	[17]
BGC0001063	Undecylprodigiosin	No	[14]
BGC0001298	Annimycin	Yes	[18]
BGC0001420	Myxochromide	Unknown	[19]
BGC0001740	Phthoxazolin	Yes	[20]

Table S4: Listing of BGCs in MIBiG that incorporates sugar monomers by glycosyltransferase in one of the tailoring steps. The ”# in BGC” and ”# in pathway” columns display the number of smCOG 1062 annotations the BGC and the number of active glycosyltransferases in the synthesis of the associated compound.

MIBiG ID	Product	# in BGC	# in pathway	Reference
BGC0000002	Aculeximycin	5	8	[21]
BGC0000021	Apoptolidin	3	2	[22]
BGC0000033	Calicheamicin	4	4	[23]
BGC0000034	Candididin	1	1	[24]
BGC0000035	Chalcomycin	2	1	[25]
BGC0000036	Chlorothricin	2	2	[4]
BGC0000042	Cremimycin	1	1	[26]
BGC0000052	ECO-02301	1	1	[27]
BGC0000054	Erythromycin B	2	2	[28]
BGC0000078	Inecidine	2	3	[29]
BGC0000081	Kedarcidin	2	2	[30]
BGC0000082	Kijanimicin	5	5	[5]
BGC0000085	Lankamycin	2	3	[31]
BGC0000092	Megalomicins	3	3	[32]
BGC0000096	Midecamycin	2	1	[33]
BGC0000102	Mycinamicin II	2	1	[34]
BGC0000105	Nanchangmycin	1	1	[35]
BGC0000108	Natamycin	1	1	[36]
BGC0000115	Nystatin A1	1	1	[37]
BGC0000136	Rifamycin	1	1	[38]
BGC0000141	Rubradirin	1	2	[39]
BGC0000148	A83543A	2	2	[40]
BGC0000151	Stambomycin A	1	1	[41]
BGC0000162	Tetrocarcin A	5	4	[7]
BGC0000165	Tiacumicin B	2	2	[42]
BGC0000167	Vicenistatin	1	1	[43]
BGC0000197	Aranciamycin	1	1	[44]
BGC0000198	Arenimycin A/B/C	2	2	[45]
BGC0000199	ArimetamycinA	2	3	[46]
BGC0000199	ArimetamycinB	1	3	[46]
BGC0000199	ArimetamycinC	1	3	[46]
BGC0000200	Arixanthomycin A	1	2	[47]
BGC0000203	BE-7585A	3	1	[48]
BGC0000208	Chelocardin	0	1	[49]
BGC0000210	Chromomycin A3	5	4	[50]
BGC0001183	Lobophorin	3	4	[9]
BGC0001452	Sipanmycin	2	4	[51]
BGC0001522	Auroramycin	2	4	[52]
BGC0001619	Ibomycin	6	7	[53]
BGC0001851	Komodoquinone B	0	5	[54]
BGC0002033	Spiramycin	3	4	[55]

Rare extender units

The smCOG gene annotations were also leveraged to determine if the genes encoding enzymes responsible for the synthesis of the rare polyketide precursor P_{ii}malonyl-CoA were present in the BGC. This synthesis pathway was identified by the presence of smCOG 1256 (FkbH like protein) and smCOG 1095 (3-hydroxybutyryl-CoA dehydrogenase), see Table S5. This heuristic predicts the correct extender unit in 16 of the 24 relevant BGCs. The relevant BGC are selected from the MIBiG database based those that contains a gene with the smCOG 1245 and incorporates one or more rare extender unit. Seven of the 8 incorrect predictions are due to BGCS incorporating hydroxymalonyl-ACP and not methoxymalonyl-ACP, and the last error derives from BGC0000090 which incorporates methoxymalonyl-ACP, but lacks the smCOG 1095 annotation. Because we are currently not able to discriminate the synthesis of hydroxymalonyl-CoA from the synthesis of methoxymalonyl-ACP based on smCOG annotations, the BiGMeC pipeline assumes that the rare extender unit is methoxymalonyl-ACP. While this determines the incorporation of the reactions synthesising methoxymalonyl-ACP [56], the incorporation of methoxymalonyl-ACP as an extender unit only occurs if this requirement is fulfilled and methoxymalonyl-ACP is the extender unit as suggested by antiSMASH [57].

Table S5: List of BGCs in MIBiG that contains a gene with the smCOG 1245 annotation and incorporates a rare extender unit.

MIBiG ID	Product	Substrate	1095 present	Reference
BGC0000020	Actinosynnema	Hydroxymalonyl-ACP	Yes	[58]
BGC0000021	Apoptolidin	Methoxymalonyl-ACP	Yes	[22]
BGC0000028	Bafilomycin	Methoxymalonyl-ACP	Yes	[15]
BGC0000040	Concanamycin A	Methoxymalonyl-ACP	Yes	[59]
BGC0000065	Rustmicin	Methoxymalonyl-ACP	Yes	[60]
BGC0000066	Geldanamycin	Methoxymalonyl-ACP	Yes	[61]
BGC0000074	Herbimycin A	Methoxymalonyl-ACP	Yes	[38]
BGC0000078	Incednine	Methoxymalonyl-ACP	Yes	[29]
BGC0000090	Macbecin	Methoxymalonyl-ACP	No	[62]
BGC0000096	Midecamycin	Methoxymalonyl-ACP	Yes	[33]
BGC0000159	Tautomycin	Methoxymalonyl-ACP	Yes	[63]
BGC0000970	Chondrochloren A	Methoxymalonyl-ACP	Yes	[64]
BGC0001034	Pellasoren	Methoxymalonyl-ACP	Yes	[65]
BGC0001054	Xenocoumacin	Hydroxymalonyl-ACP	Yes	[66]
BGC0001059	Zwittermycin A	Hydroxymalonyl-ACP	Yes	[67]
BGC0001106	Oxazolomycin B	Methoxymalonyl-ACP	Yes	[68]
BGC0001348	JBIR-100	Methoxymalonyl-ACP	Yes	[69]
BGC0001511	Ansamitocin P-3	Methoxymalonyl-ACP	Yes	[70]
BGC0001537	Butyrolactol A	Hydroxymalonyl-ACP	Yes	[71]
BGC0001902	Bengamide	Hydroxymalonyl-ACP	Yes	[72]
BGC0001956	Miharamycin A	Hydroxymalonyl-ACP	Yes	[73]
BGC0001957	Amipurimycin	Hydroxymalonyl-ACP	Yes	[73]
BGC0002011	Ansacarbamitocin A	Methoxymalonyl-ACP	Yes	[74]
BGC0002033	Spiramycin	Methoxymalonyl-ACP	Yes	[55]

References

- [1] Kautsar, S.A., Blin, K., Shaw, S., Navarro-Muñoz, J.C., Terlouw, B.R., van der Hoof, J.J., Van Santen, J.A., Tracanna, V., Suarez Duran, H.G., Pascal Andreu, V., *et al.*: Mibig 2.0: a repository for biosynthetic gene clusters of known function. *Nucleic acids research* **48**(D1), 454–458 (2020)
- [2] Medema, M.H., Blin, K., Cimermanic, P., de Jager, V., Zakrzewski, P., Fischbach, M.A., Weber, T., Takano, E., Breitling, R.: antimash: rapid identification, annotation and analysis of secondary metabolite biosynthesis gene clusters in bacterial and fungal genome sequences. *Nucleic acids research* **39**(suppl.2), 339–346 (2011)
- [3] Gottardi, E.M., Krawczyk, J.M., von Suchodoletz, H., Schadt, S., Mühlenweg, A., Uguru, G.C., Pelzer, S., Fiedler, H.-P., Bibb, M.J., Stach, J.E.M., Süßmuth, R.D.: Abyssomicin biosynthesis: formation of an unusual polyketide, antibiotic-feeding studies and genetic analysis. *Chembiochem : a European journal of chemical biology* **12**(9), 1401–1410 (2011)
- [4] Jia, X.-Y., Tian, Z.-H., Shao, L., Qu, X.-D., Zhao, Q.-F., Tang, J., Tang, G.-L., Liu, W.: Genetic characterization of the chlorothricin gene cluster as a model for spirotetronate antibiotic biosynthesis. *Chemistry & Biology* **13**(6), 575–585 (2006)
- [5] Zhang, H., White-Phillip, J.A., Melancon, r. Charles E, Kwon, H.-j., Yu, W.-l., Liu, H.-w.: Elucidation of the kijanimicin gene cluster: insights into the biosynthesis of spirotetronate antibiotics and nitrosugars. *Journal of the American Chemical Society* **129**(47), 14670–14683 (2007)
- [6] He, H.-Y., Pan, H.-X., Wu, L.-F., Zhang, B.-B., Chai, H.-B., Liu, W., Tang, G.-L.: Quartromicin biosynthesis: Two alternative polyketide chains produced by one polyketide synthase assembly line. *Chemistry & Biology* **19**(10), 1313–1323 (2012)
- [7] Fang, J., Zhang, Y., Huang, L., Jia, X., Zhang, Q., Zhang, X., Tang, G., Liu, W.: Cloning and characterization of the tetrocarcin a gene cluster from *micromonospora chalcea* nrrl 11289 reveals a highly conserved strategy for tetronate biosynthesis in spirotetronate antibiotics. *Journal of Bacteriology* **190**(17), 6014–6025 (2008)
- [8] Demydchuk, Y., Sun, Y., Hong, H., Staunton, J., Spencer, J.B., Leadlay, P.F.: Analysis of the tetronomycin gene cluster: Insights into the biosynthesis of a polyether tetronate antibiotic. *ChemBioChem* **9**(7), 1136–1145 (2008)

- [9] Zhang, C., Ding, W., Qin, X., Ju, J.: Genome sequencing of *Streptomyces olivaceus* scsio t05 and activated production of lobophorin cr4 via metabolic engineering and genome mining. *Marine drugs* **17**(10), 593 (2019)
- [10] Hashimoto, T., Hashimoto, J., Teruya, K., Hirano, T., Shin-ya, K., Ikeda, H., Liu, H.-w., Nishiyama, M., Kuzuyama, T.: Biosynthesis of versipelostatin: Identification of an enzyme-catalyzed [4+2]-cycloaddition required for macrocyclization of spirotetronate-containing polyketides. *Journal of the American Chemical Society* **137**(2), 572–575 (2015)
- [11] Daduang, R., Kitani, S., Hashimoto, J., Thamchaipenet, A., Igarashi, Y., Shin-ya, K., Ikeda, H., Nihira, T.: Characterization of the biosynthetic gene cluster for maklamicin, a spirotetronate-class antibiotic of the endophytic micromonospora sp. NBRC 110955. *Microbiological Research* **180**, 30–39 (2015)
- [12] Zhang, W., Bolla, M.L., Kahne, D., Walsh, C.T.: A three enzyme pathway for 2-amino-3-hydroxycyclopent-2-enone formation and incorporation in natural product biosynthesis. *Journal of the American Chemical Society* **132**(18), 6402–6411 (2010)
- [13] Salem, S.M., Kancharla, P., Florova, G., Gupta, S., Lu, W., Reynolds, K.A.: Elucidation of final steps of the marineosins biosynthetic pathway through identification and characterization of the corresponding gene cluster. *Journal of the American Chemical Society* **136**(12), 4565–4574 (2014)
- [14] Williamson, N.R., Simonsen, H.T., Ahmed, R.A.A., Goldet, G., Slater, H., Woodley, L., Leeper, F.J., Salmond, G.P.C.: Biosynthesis of the red antibiotic, prodigiosin, in *Serratia*: identification of a novel 2-methyl-3-n-amylopyrrole (map) assembly pathway, definition of the terminal condensing enzyme, and implications for undecylprodigiosin biosynthesis in *Streptomyces*. *Molecular Microbiology* **56**(4), 971–989 (2005)
- [15] Nara, A., Hashimoto, T., Komatsu, M., Nishiyama, M., Kuzuyama, T., Ikeda, H.: Characterization of bafilomycin biosynthesis in *Kitasatospora setae* km-6054 and comparative analysis of gene clusters in actinomycetales microorganisms. *The Journal of Antibiotics* **70**(5), 616–624 (2017)
- [16] Rui, Z., Petřicková, K., Skanta, F., Pospíšil, S., Yang, Y., Chen, C.-Y., Tsai, S.-F., Floss, H.G., Petřicek, M., Yu, T.-W.: Biochemical and genetic insights into asukamycin biosynthesis. *The Journal of biological chemistry* **285**(32), 24915–24924 (2010)

- [17] Petříčková, K., Pospíšil, S., Kuzma, M., Tylová, T., Jágr, M., Tomek, P., Chroňáková, A., Brabcová, E., Anděra, L., Kristůfek, V., Petříček, M.: Biosynthesis of colabomycin e, a new manumycin-family metabolite, involves an unusual chain-length factor. *ChemBioChem* **15**(9), 1334–1345 (2014)
- [18] Kalan, L., Gessner, A., Thaker, M.N., Waglechner, N., Zhu, X., Szawiola, A., Bechthold, A., Wright, G.D., Zechel, D.L.: A cryptic polyene biosynthetic gene cluster in *Streptomyces calvus* is expressed upon complementation with a functional bldA gene. *Chemistry & Biology* **20**(10), 1214–1224 (2013)
- [19] Burgard, C., Zaburanyi, N., Nadmid, S., Maier, J., Jenke-Kodama, H., Luxenburger, E., Bernauer, H.S., Wenzel, S.C.: Genomics-guided exploitation of lipopeptide diversity in myxobacteria. *ACS Chemical Biology* **12**(3), 779–786 (2017)
- [20] Suroto, D.A., Kitani, S., Arai, M., Ikeda, H., Nihira, T.: Characterization of the biosynthetic gene cluster for cryptic phthoxazolin a in *streptomyces avermitilis*. *PLOS ONE* **13**(1), 1–17 (2018)
- [21] Rebets, Y., Tokovenko, B., Lushchik, I., Rückert, C., Zaburanyi, N., Bechthold, A., Kalinowski, J., Luzhetskyy, A.: Complete genome sequence of producer of the glycopeptide antibiotic aculeximycin *kutzneria albida* dsm 43870t, a representative of minor genus of pseudonocardiaaceae. *BMC Genomics* **15**(1), 885 (2014)
- [22] Du, Y., Derewacz, D.K., Deguire, S.M., Teske, J., Ravel, J., Sulikowski, G.A., Bachmann, B.O.: Biosynthesis of the apoptolidins in *Nocardiopepsis* sp. fu 40. *Tetrahedron* **67**(35), 6568–6575 (2011)
- [23] Zhang, C., Bitto, E., Goff, R.D., Singh, S., Bingman, C.A., Griffith, B.R., Albermann, C., Phillips Jr, G.N., Thorson, J.S.: Biochemical and structural insights of the early glycosylation steps in calicheamicin biosynthesis. *Chemistry & biology* **15**(8), 842–853 (2008)
- [24] Gil, J., Campelo-Diez, A.: Candicidin biosynthesis in *streptomyces griseus*. *Applied Microbiology and Biotechnology* **60**(6), 633–642 (2003)
- [25] Ward, S.L., Hu, Z., Schirmer, A., Reid, R., Revill, W.P., Reeves, C.D., Petrakovsky, O.V., Dong, S.D., Katz, L.: Chalcomycin biosynthesis gene cluster from *Streptomyces bikiniensis*: Novel features of an unusual ketolide produced through expression of the chm polyketide synthase in *streptomyces fradiae*. *Antimicrobial Agents and Chemotherapy* **48**(12), 4703–4712 (2004)

- [26] Amagai, K., Takaku, R., Kudo, F., Eguchi, T.: A unique amino transfer mechanism for constructing the β -amino fatty acid starter unit in the biosynthesis of the macrolactam antibiotic cremimycin. *ChemBioChem* **14**(15), 1998–2006 (2013)
- [27] Zhang, W., Bolla, M.L., Kahne, D., Walsh, C.T.: A three enzyme pathway for 2-amino-3-hydroxycyclopent-2-enone formation and incorporation in natural product biosynthesis. *Journal of the American Chemical Society* **132**(18), 6402–6411 (2010)
- [28] Zhang, H., Wang, Y., Wu, J., Skalina, K., Pfeifer, B.A.: Complete biosynthesis of erythromycin a and designed analogs using *e. coli* as a heterologous host. *Chemistry & Biology* **17**(11), 1232–1240 (2010)
- [29] Takaishi, M., Kudo, F., Eguchi, T.: Biosynthetic pathway of 24-membered macrolactam glycoside incednine. *Tetrahedron* **64**(28), 6651–6656 (2008)
- [30] Lohman, J.R., Huang, S.-X., Horsman, G.P., Dilfer, P.E., Huang, T., Chen, Y., Wendt-Pienkowski, E., Shen, B.: Cloning and sequencing of the kedarcidin biosynthetic gene cluster from *streptoalloteichus* sp. atcc 53650 revealing new insights into biosynthesis of the enediyne family of antitumor antibiotics. *Molecular bioSystems* **9**(3), 478–491 (2013)
- [31] Arakawa, K., Kodama, K., Tatsuno, S., Ide, S., Kinashi, H.: Analysis of the loading and hydroxylation steps in lankamycin biosynthesis in *streptomyces rochei*. *Antimicrobial agents and chemotherapy* **50**(6), 1946–1952 (2006)
- [32] Volchegursky, Y., Hu, Z., Katz, L., McDaniel, R.: Biosynthesis of the anti-parasitic agent megalomicin: transformation of erythromycin to megalomicin in *saccharopolyspora erythraea*. *Molecular Microbiology* **37**(4), 752–762 (2000)
- [33] Cong, L., Piepersberg, W.: Cloning and Characterization of Genes Encoded in dTDP-D-mycaminose Biosynthetic Pathway from a Midecamycin-producing Strain, *Streptomyces mycarofaciens*. *Acta Biochimica et Biophysica Sinica* **39**(3), 187–193 (2007)
- [34] Anzai, Y., Tsukada, S.-i., Sakai, A., Masuda, R., Harada, C., Domeki, A., Li, S., Kinoshita, K., Sherman, D., Kato, F.: Function of cytochrome p450 enzymes mycc1 and mycg in *micromonospora griseorubida*, a producer of the macrolide antibiotic mycinamicin. *Antimicrobial agents and chemotherapy* **56**, 3648–56 (2012)
- [35] Liu, T., Lin, X., Zhou, X., Deng, Z., Cane, D.E.: Mechanism of thioesterase-catalyzed chain release in the biosynthesis of the polyether antibiotic nanchangmycin. *Chemistry & biology* **15**(5), 449–458 (2008)

- [36] Aparicio, J.F., Barreales, E.G., Payero, T.D., Vicente, C.M., de Pedro, A., Santos-Aberturas, J.: Biotechnological production and application of the antibiotic pimaricin: biosynthesis and its regulation. *Applied microbiology and biotechnology* **100**(1), 61–78 (2016)
- [37] Bruheim, P., Borgos, S.E., Tsan, P., Sletta, H., Ellingsen, T., Lancelin, J.-M., Zotchev, S.: Chemical diversity of polyene macrolides produced by *Streptomyces noursei* ATCC 11455 and recombinant strain Erd44 with genetically altered polyketide synthase. *Antimicrobial Agents and Chemotherapy* **48**, 4120–9 (2004)
- [38] Chapter 13 - alkaloids derived from an m-c7n unit. In: Funayama, S., Cordell, G.A. (eds.) *Alkaloids*, pp. 219–232. Academic Press, Boston (2015)
- [39] Kim, C.-G., Lamichhane, J., Song, K.-I., Nguyen, V.D., Kim, D.-H., Jeong, T.-S., Kang, S.-H., Kim, K.-W., Maharjan, J., Hong, Y.-S., Kang, J.S., Yoo, J.-C., Lee, J.-J., Oh, T.-J., Liou, K., Sohng, J.K.: Biosynthesis of rubradirin as an ansamycin antibiotic from *Streptomyces achromogenes* var. *rubradiris* NRRL3061. *Archives of Microbiology* **189**(5), 463–473 (2007)
- [40] Evans, D.A., Black, W.C.: Total synthesis of (+)-a83543a [(+)-lepicidin a]. *Journal of the American Chemical Society* **115**(11), 4497–4513 (1993)
- [41] Song, L., Laureti, L., Corre, C., Leblond, P., Aigle, B., Challis, G.L.: Cytochrome p450-mediated hydroxylation is required for polyketide macrolactonization in stambomycin biosynthesis. *The Journal of Antibiotics* **67**(1), 71–76 (2014)
- [42] Xiao, Y., Li, S., Niu, S., Ma, L., Zhang, G., Zhang, H., Zhang, G., Ju, J., Zhang, C.: Characterization of tiacumicin B biosynthetic gene cluster affording diversified tiacumicin analogues and revealing a tailoring dihalogenase. *Journal of the American Chemical Society* **133**(4), 1092–1105 (2011)
- [43] Ogasawara, Y., Katayama, K., Minami, A., Otsuka, M., Eguchi, T., Kakinuma, K.: Cloning, sequencing, and functional analysis of the biosynthetic gene cluster of macrolactam antibiotic vicenistatin in *Streptomyces halstedii*. *Chemistry & Biology* **11**(1), 79–86 (2004)
- [44] Luzhetskyy, A., Mayer, A., Hoffmann, J., Pelzer, S., Holzenkämper, M., Schmitt, B., Wohlert, S.-E., Vente, A., Bechthold, A.: Cloning and heterologous expression of the aranciamycin biosynthetic gene cluster revealed a new flexible glycosyltransferase. *ChemBioChem* **8**(6), 599–602 (2007)

- [45] Jensen, P.R., Moore, B.S., Fenical, W.: The marine actinomycete genus *salinispora*: a model organism for secondary metabolite discovery. *Natural product reports* **32**(5), 738–751 (2015)
- [46] Kang, H.-S., Brady, S.F.: Arimetamycin a: Improving clinically relevant families of natural products through sequence-guided screening of soil metagenomes. *Angewandte Chemie International Edition* **52**(42), 11063–11067 (2013)
- [47] Kang, H.-S., Brady, S.F.: Arixanthomycins a-c: Phylogeny-guided discovery of biologically active edna-derived pentangular polyphenols. *ACS chemical biology* **9**(6), 1267–1272 (2014)
- [48] Sasaki, E., Ogasawara, Y., Liu, H.-w.: A biosynthetic pathway for be-7585a, a 2-thiosugar-containing angucycline-type natural product. *Journal of the American Chemical Society* **132**(21), 7405–7417 (2010)
- [49] Lukežič, T., Lešnik, U., Podgoršek, A., Horvat, J., Polak, T., Šala, M., Jenko, B., Raspor, P., Herron, P.R., Hunter, I.S., *et al.*: Identification of the chelocardin biosynthetic gene cluster from *amycolatopsis sulphurea*: a platform for producing novel tetracycline antibiotics. *Microbiology* **159**(Pt.12), 2524–2532 (2013)
- [50] Menéndez, N., Nur-e-Alam, M., Braña, A.F., Rohr, J., Salas, J.A., Méndez, C.: Biosynthesis of the antitumor chromomycin a3 in *Streptomyces griseus*: Analysis of the gene cluster and rational design of novel chromomycin analogs. *Chemistry & Biology* **11**(1), 21–32 (2004)
- [51] Malmierca, M.G., Pérez-Victoria, I., Martín, J., Reyes, F., Méndez, C., Salas, J.A., Olano, C.: New sipanmycin analogues generated by combinatorial biosynthesis and mutasynthesis approaches relying on the substrate flexibility of key enzymes in the biosynthetic pathway. *Applied and Environmental Microbiology* **86**(3) (2020)
- [52] Yeo, W.L., Heng, E., Tan, L.L., Lim, Y.W., Ching, K.C., Tsai, D.-J., Jhang, Y.W., Lauderdale, T.-L., Shia, K.-S., Zhao, H., Ang, E.L., Zhang, M.M., Lim, Y.H., Wong, F.T.: Biosynthetic engineering of the antifungal, anti-mrsa auroramycin. *Microbial Cell Factories* **19**(1), 3 (2020)
- [53] Robbins, N., Spitzer, M., Wang, W., Waglechner, N., Patel, D.J., O’Brien, J.S., Ejim, L., Ejim, O., Tyers, M., Wright, G.D.: Discovery of ibomycin, a complex macrolactone that exerts antifungal activity by impeding endocytic trafficking and membrane function. *Cell Chemical Biology* **23**(11), 1383–1394 (2016)

- [54] Grocholski, T., Yamada, K., Sinkkonen, J., Tirkkonen, H., Niemi, J., Metsä-Ketelä, M.: Evolutionary trajectories for the functional diversification of anthracycline methyltransferases. *ACS chemical biology* **14**(5), 850–856 (2019)
- [55] Karray, F., Darbon, E., Oestreicher, N., Dominguez, H., Tuphile, K., Gagnat, J., Blondelet-Rouault, M.-H., Gerbaud, C., Pernodet, J.-L.: Organization of the biosynthetic gene cluster for the macrolide antibiotic spiramycin in *Streptomyces ambofaciens*. *Microbiology* **153**(12), 4111–4122 (2007)
- [56] Chan, Y.A., Podevels, A.M., Kevany, B.M., Thomas, M.G.: Biosynthesis of polyketide synthase extender units. *Natural product reports* **26**(1), 90–114 (2009)
- [57] Blin, K., Shaw, S., Steinke, K., Villebro, R., Ziemert, N., Lee, S.Y., Medema, M.H., Weber, T.: antimash 5.0: updates to the secondary metabolite genome mining pipeline. *Nucleic acids research* **47**(W1), 81–87 (2019)
- [58] Yu, T.-W., Bai, L., Clade, D., Hoffmann, D., Toelzer, S., Trinh, K.Q., Xu, J., Moss, S.J., Leistner, E., Floss, H.G.: The biosynthetic gene cluster of the maytansinoid antitumor agent ansamitocin from *Actinosynnema pretiosum*. *Proceedings of the National Academy of Sciences* **99**(12), 7968–7973 (2002)
- [59] Haydock, S.F., Appleyard, A.N., Mironenko, T., Lester, J., Scott, N., Leadlay, P.F.: Organization of the biosynthetic gene cluster for the macrolide concanamycin A in *Streptomyces neyagawaensis* atcc 27449. *Microbiology* **151**(10), 3161–3169 (2005)
- [60] Mandala, S.M., Thornton, R.A., Milligan, J., Rosenbach, M., Garcia-Calvo, M., Bull, H.G., Harris, G., Abruzzo, G.K., Flattery, A.M., Gill, C.J., Bartizal, K., Dreikorn, S., Kurtz, M.B.: Rustmicin, a potent antifungal agent, inhibits sphingolipid synthesis at inositol phosphoceramide synthase. *Journal of Biological Chemistry* **273**(24), 14942–14949 (1998)
- [61] Patel, K., Piagentini, M., Rascher, A., Tian, Z.-Q., Buchanan, G.O., Regentin, R., Hu, Z., Hutchinson, C.R., McDaniel, R.: Engineered biosynthesis of geldanamycin analogs for hsp90 inhibition. *Chemistry & Biology* **11**(12), 1625–1633 (2004)
- [62] Hatano, K., Muroi, M., Higashide, E., Yoneda, M.: Biosynthesis of macbecin. *Agricultural and Biological Chemistry* **46**(6), 1699–1702 (1982)

- [63] Li, W., Ju, J., Rajski, S.R., Osada, H., Shen, B.: Characterization of the tautomycin biosynthetic gene cluster from *Streptomyces spiroverticillatus* Unveiling new insights into dialkylmaleic anhydride and polyketide biosynthesis. *Journal of Biological Chemistry* **283**(42), 28607–28617 (2008)
- [64] Rachid, S., Scharfe, M., Blöcker, H., Weissman, K.J., Müller, R.: Unusual chemistry in the biosynthesis of the antibiotic chondrochlorens. *Chemistry & Biology* **16**(1), 70–81 (2009)
- [65] Jahns, C., Hoffmann, T., Müller, S., Gerth, K., Washausen, P., Höfle, G., Reichenbach, H., Kalesse, M., Müller, R.: Pellasoren: Structure elucidation, biosynthesis, and total synthesis of a cytotoxic secondary metabolite from *Sorangium cellulosum*. *Angewandte Chemie International Edition* **51**(21), 5239–5243 (2012)
- [66] Masschelein, J., Jenner, M., Challis, G.: Antibiotics from gram-negative bacteria: A comprehensive overview and selected biosynthetic highlights. *Natural Product Reports* **34** (2017)
- [67] Kevany, B.M., Rasko, D.A., Thomas, M.G.: Characterization of the complete zwittermicin a biosynthesis gene cluster from *Bacillus cereus*. *Applied and Environmental Microbiology* **75**(4), 1144–1155 (2009)
- [68] Helfrich, E.J.N., Piel, J.: Biosynthesis of polyketides by trans-AT polyketide synthases. *Natural Product Reports* **33**(2), 231–316 (2016)
- [69] Molloy, E.M., Tietz, J.I., Blair, P.M., Mitchell, D.A.: Biological characterization of the hygrobafilomycin antibiotic jbir-100 and bioinformatic insights into the hygrolide family of natural products. *Bioorganic & medicinal chemistry* **24**(24), 6276–6290 (2016)
- [70] Lin, J., Bai, L., Deng, Z., Zhong, J.-J.: Effect of ammonium in medium on ansamitocin p-3 production by *Actinosynnema pretiosum*. *Biotechnology and Bioprocess Engineering* **15**, 119–125 (2010)
- [71] Harunari, E., Komaki, H., Igarashi, Y.: Biosynthetic origin of butyrolactol a, an antifungal polyketide produced by a marine-derived streptomyces. *Beilstein journal of organic chemistry* **13**, 441–450 (2017)
- [72] Wenzel, S.C., Hoffmann, H., Zhang, J., Debussche, L., Haag-Richter, S., Kurz, M., Nardi, F., Lukat, P., Kochems, I., Tietgen, H., Schummer, D., Nicolas, J.-P., Calvet, L., Czepczor, V., Vrignaud, P., Mühlenweg, A., Pelzer, S., Müller, R., Brönstrup, M.: Production of the bengamide class of marine natural products in myxobacteria: Biosynthesis and structure–activity relationships. *Angewandte Chemie International Edition* **54**(51), 15560–15564 (2015)

- [73] Romo, A.J., Shiraishi, T., Ikeuchi, H., Lin, G.-M., Geng, Y., Lee, Y.-H., Liem, P.H., Ma, T., Ogasawara, Y., Shin-ya, K., Nishiyama, M., Kuzuyama, T., Liu, H.-w.: The amipurimycin and miharamycin biosynthetic gene clusters: Unraveling the origins of 2-aminopurinylyl peptidyl nucleoside antibiotics. *Journal of the American Chemical Society* **141**(36), 14152–14159 (2019)
- [74] Li, X., Wu, X., Shen, Y.: Identification of the bacterial maytansinoid gene cluster *asc* provides insights into the post-pks modifications of ansacarbamitocin biosynthesis. *Organic Letters* **21**(15), 5823–5826 (2019)

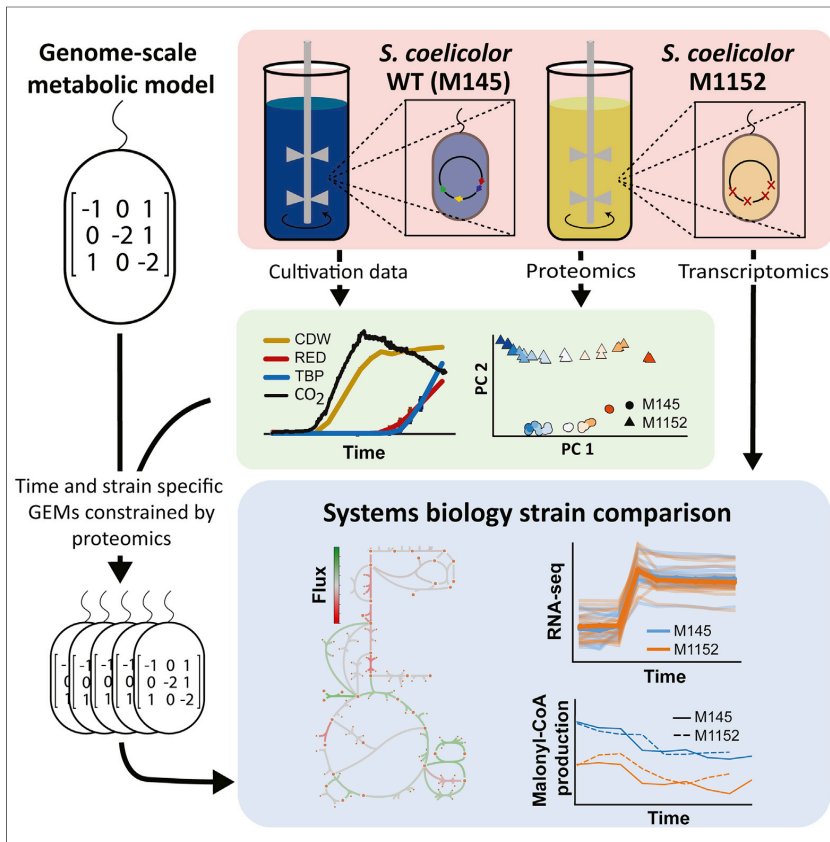
Paper 4

Enzyme-Constrained Models and Omics Analysis of *Streptomyces coelicolor* Reveal Metabolic Changes that Enhance Heterologous Production

Snorre Sulheim, Tjaša Kumelj, Dino van Dissel, Ali Salehzadeh-Yazdi, Chao Du, Gilles P. van Wezel, Kay Nieselt, Alexander Wentzel, Eivind Almaas and Eduard Kerkhoven. *iScience*, 23(9), 101525 (2020).

Article

Enzyme-Constrained Models and Omics Analysis of *Streptomyces coelicolor* Reveal Metabolic Changes that Enhance Heterologous Production



Snorre Sulheim, Tjaša Kumelj, Dino van Dissel, ..., Eivind Almaas, Alexander Wentzel, Eduard J. Kerkhoven

eduardk@chalmers.se

HIGHLIGHTS

Time-series transcriptomics and proteomics of *S. coelicolor* M145 and M1152

Application of GEM to interpret changes in the proteome on the systems level

Limited effect of improved precursor supply on enhanced production in M1152

Reduced rate of germicidin in M1152 suggests a need for other expression hosts

Sulheim et al., iScience 23, 101525
September 25, 2020 © 2020 The Authors.
<https://doi.org/10.1016/j.isci.2020.101525>



Article

Enzyme-Constrained Models and Omics Analysis of *Streptomyces coelicolor* Reveal Metabolic Changes that Enhance Heterologous Production

Snorre Sulheim,^{1,2} Tjaša Kumelj,² Dino van Dissel,¹ Ali Salehzadeh-Yazdi,³ Chao Du,⁴ Gilles P. van Wezel,⁴ Kay Nieselt,⁵ Eivind Almaas,^{2,6} Alexander Wentzel,¹ and Eduard J. Kerkhoven^{7,8,9,*}

SUMMARY

Many biosynthetic gene clusters (BGCs) require heterologous expression to realize their genetic potential, including silent and metagenomic BGCs. Although the engineered *Streptomyces coelicolor* M1152 is a widely used host for heterologous expression of BGCs, a systemic understanding of how its genetic modifications affect the metabolism is lacking and limiting further development. We performed a comparative analysis of M1152 and its ancestor M145, connecting information from proteomics, transcriptomics, and cultivation data into a comprehensive picture of the metabolic differences between these strains. Instrumental to this comparison was the application of an improved consensus genome-scale metabolic model (GEM) of *S. coelicolor*. Although many metabolic patterns are retained in M1152, we find that this strain suffers from oxidative stress, possibly caused by increased oxidative metabolism. Furthermore, precursor availability is likely not limiting polyketide production, implying that other strategies could be beneficial for further development of *S. coelicolor* for heterologous production of novel compounds.

INTRODUCTION

The bacterium *Streptomyces coelicolor* has been the *de facto* model actinomycete for the production of antibiotics. Being known for over 100 years, the interest in this organism predates the golden age of antibiotic research. With its complex life cycle, featuring mycelial growth and differentiation, spore formation, programmed cell death, and the ability to produce multiple colored secondary metabolites, it has assisted greatly in our understanding of how streptomycetes sense their surrounding (Hahn et al., 2002; Hutchings et al., 2004; Nothaft et al., 2010; Rigali et al., 2008; Sola-Landa et al., 2005), activate their developmental cycle (Chandra and Chater, 2014), and regulate the production of antibiotics (Nieselt et al., 2010; Thomas et al., 2012). Further aided by the publication of its genome sequence (Bentley et al., 2002), the antibiotic coelimycin P1 (yellow), produced from the formerly cryptic polyketide gene cluster known as *cpk*, was added to this list (Gomez-Escribano et al., 2012). Today, the widespread use of *S. coelicolor* continues as a host for heterologous production of biosynthetic gene clusters (BGCs) (Castro et al., 2015; Gomez-Escribano and Bibb, 2011, 2014; Kumelj et al., 2019; Thanapipatsiri et al., 2015; Yin et al., 2015). Heterologous expression is a powerful strategy for novel compound discovery from BGCs that are either natively silent or originate from an unculturable source (Nepal and Wang, 2019). These BGCs represent an untapped resource of microbial biodiversity, nowadays made evident and accessible due to recent advances within the fields of metagenomics, molecular biology, and bioinformatics (Rutledge and Challis, 2015).

The efficiency of *S. coelicolor* as a heterologous production host relies on a metabolism that has evolved to provide the necessary precursors to produce a broad range of complex molecules. Many of these molecules are produced when the strain is experiencing nutrient-limiting conditions that lead to growth cessation and complex re-modelling of its metabolism (Wentzel et al., 2012a). Metabolic switching in response to phosphate and glutamate depletion has been studied in detail at a variety of metabolic levels in *S. coelicolor* M145 (Nieselt et al., 2010; Thomas et al., 2012; Wentzel et al., 2012b), the most well-known wild-type strain devoid of the two plasmids SCP1 and SCP2 present in the parent strain *S. coelicolor* A3(2) (Kieser et al., 2000). This has unraveled a complex sequence of switching events that ultimately lead to the biosynthesis of calcium-dependent antibiotic (CDA), and the colored antibiotics actinorhodin

¹Department of Biotechnology and Nanomedicine, SINTEF Industry, 7034 Trondheim, Norway

²Department of Biotechnology and Food Science, NTNU - Norwegian University of Science and Technology, 7491 Trondheim, Norway

³Department of Systems Biology and Bioinformatics, Faculty of Computer Science and Electrical Engineering, University of Rostock, 18057 Rostock, Germany

⁴Microbial Biotechnology, Institute of Biology, Leiden University, 2300 Leiden, the Netherlands

⁵Integrative Transcriptomics, Center for Bioinformatics, University of Tübingen, 72070 Tübingen, Germany

⁶K.G. Jebsen Center for Genetic Epidemiology, Department of Public Health and General Practice, NTNU - Norwegian University of Science and Technology, 7491 Trondheim, Norway

⁷Systems and Synthetic Biology, Department of Biology and Biological Engineering, Chalmers University of Technology, 412 96 Gothenburg, Sweden

⁸Novo Nordisk Foundation Center for Biosustainability, Chalmers University of Technology, 412 96 Gothenburg, Sweden

⁹Lead Contact

*Correspondence: eduardk@chalmers.se

<https://doi.org/10.1016/j.isci.2020.101525>



(Act, blue) and undecylprodigiosin (Red, red). The biosynthesis of coelimycin P1 occurs earlier than the three other compounds in the growth cycle and appears to be independent of the major metabolic switch (Nieselt et al., 2010).

To improve *S. coelicolor* M145 as a host for heterologous BGC expression, strain M1146 was created by the sequential deletion of its four major BGCs (*act*, *red*, *cda*, and *cpk*) (Gomez-Escribano and Bibb, 2011). This should increase precursor availability for the production of a whole range of heterologous products and provides a cleaner chromatographic background to more easily identify novel compounds. *S. coelicolor* M1152 is a derivative of M1146, which besides the deletion of the four main BGCs bears the C1298T point mutation in the *rpoB* gene that encodes the beta subunit of RNA polymerase. This mutation was shown to have strong positive effects on the production of various antibiotics (Gomez-Escribano and Bibb, 2011; Hu et al., 2002). Up to now, M1152 is a preferred general “superhost” for heterologous BGC expression (Braesel et al., 2019; Castro et al., 2015; Kepplinger et al., 2018; Li et al., 2013; Thanapipatsiri et al., 2015) and is the starting point for further strain development.

Previous research on the metabolism of *S. coelicolor* M1152 has been confined to transcriptome profiling of batch fermentations (Battke et al., 2010; Jager et al., 2011; Liao et al., 2014; Love et al., 2014; Mi et al., 2019), and further development of this strain as a “superhost” calls for a better understanding of how the genetic modifications have affected the regulatory system and metabolism of M1152. To this end we measure both protein and transcript levels of both M1152 and its parent strain, M145, at different time steps during batch fermentation where the metabolic switch is triggered by depletion of phosphate. As enzymes are catalyzing most metabolic transformations, assessing protein abundance provides information about the metabolic capacity of the organism. Furthermore, we do not only consider the protein abundances in isolation but also use these measurements to confine fluxes predicted by a genome-scale metabolic model (GEM) of *S. coelicolor* to the maximum capacity of the enzymes. By doing so we propagate differences in the abundance of individual enzymes in M145 and M1152 to metabolic rearrangements on the systems level.

The metabolic network in the cell is described in a GEM (Gu et al., 2019). GEMs are valuable resources of strain-specific knowledge, mathematical models able to predict steady-state flux distributions, and frameworks for interpretation and integration of different “omics” data, e.g., transcriptomics and proteomics (Robinson and Nielsen, 2016). The increased interest in using genome-scale models of *S. coelicolor* is conspicuous. Since the first reconstruction in 2005 (Borodina et al., 2005) five GEMs have been published (Alam et al., 2010; Amara et al., 2018; Kim et al., 2014; Kumelj et al., 2019; Wang et al., 2018), including three in 2018: iKS1317 (Kumelj et al., 2019), Sco4 (Wang et al., 2018), and iAA1259 (Amara et al., 2018). In addition, as a model organism for the Actinomycetes, the GEMs of *S. coelicolor* are frequently used as template for model development of closely related strains (Mohite et al., 2019), such as *Streptomyces clavuligerus* (Toro et al., 2018), *Saccharopolyspora erythraea* (Licona-Cassani et al., 2012) and *Streptomyces lividans* (Valverde et al., 2018). The recent updates of the *S. coelicolor* GEM were developed in parallel by different research groups: although all groups share the common interest of utilizing a high-quality model for predictions and data analysis, the prevailing approach of independent parallel development is inefficient. In addition to duplicating a considerable amount of work, lack of common standards for documentation of progress and issues, evaluation of model performance, as well as the use of different annotations makes it cumbersome to compare and merge models.

To increase the rate and quality of model reconstruction, in this study two research groups of the *S. coelicolor* GEM community, responsible for two of the latest model updates (Kumelj et al., 2019; Wang et al., 2018), have joined forces to merge existing GEMs of *S. coelicolor* into one consensus model that is publicly hosted on GitHub and can be continuously updated and improved by all members of the community. Hosting the model on GitHub has many advantages: (1) open access and contribution, (2) version control, (3) continuous development and integrated quality control with memote (Lieven et al., 2020), (4) new improvements released instantly (no publication lag time), and (5) complete documentation of model reconstruction. Such an approach has historic precedents: model reconstruction as a community effort has been a success for the human GEM (Thiele et al., 2013), baker’s yeast (Aung et al., 2013; Dobson et al., 2010; Heavner et al., 2012, 2013; Herrgard et al., 2008; Lu et al., 2019), and Chinese hamster ovary cells (Hefzi et al., 2016). The recent developments in *S. coelicolor* model and strain improvements in different research groups prove that it is an opportune time now to join forces in the *Streptomyces* modeling efforts as well.

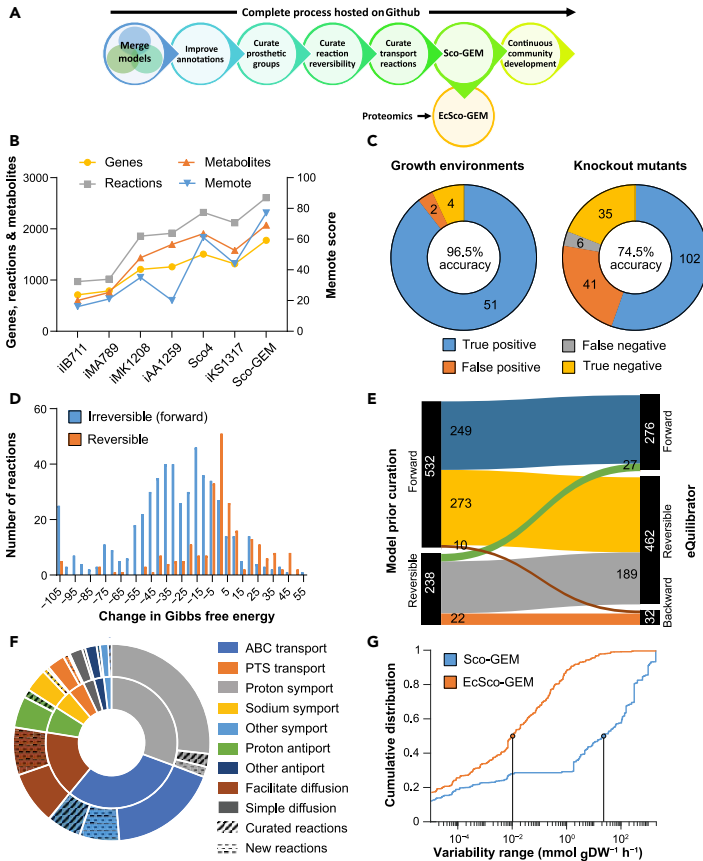


Figure 1. Sco-GEM Development and Analysis

(A) Schematic overview of the various steps in the Sco-GEM reconstruction process.

(B) The overall memote score and number of genes, reactions, and metabolites for the seven published *S. coelicolor* GEMs.

(C) Assessment of the model quality by comparing *in vivo* observations with *in silico* predictions across in total 241 tests: accuracy = 0.80; sensitivity = 0.96; specificity = 0.48; Matthews correlation coefficient = 0.53.

(D) The change in Gibbs free energy for 770 reactions that were annotated as either reversible or forward (i.e., forward irreversible) in the model before curation of reaction reversibility. The histogram is truncated at -105 kJ/mol, and more negative values are assigned to the leftmost bin.

(E) Analysis and comparison of the directionality and reversibility of reactions before curation and the direction inferred from the change in Gibbs free energy as estimated by eQuilibrator. Reactions labeled “forward” or “backward” are irreversible.

(F) Overview of the 369 transport reactions included in Sco-GEM, whereof 42 were curated and 65 were added during this work. The inner ring categorizes the reactions into nine different subgroups, whereas the outer ring displays the amount of curated and added reactions within each category. In the outer ring, the sections representing curated and new reactions are hatched and dotted, respectively.

(G) Comparison of cumulative flux variability distributions in Sco-GEM and EcSco-GEM.

RESULTS

Reconstruction of the Consensus Genome-Scale Model of *S. Coelicolor*

We conducted a stepwise reconstruction of Sco-GEM, the consensus genome-scale metabolic model of *S. coelicolor*, while tracking development using Git for version control (Figure 1A, Data S1, Table 1).

SCO-GEM is the most comprehensive and highest quality GEM of this organism (Figure 1B), comprising 1,777 genes, 2,612 reactions, 2,073 metabolites, and a memote score of 77%, which is indicative of the overall model quality (Lieven et al., 2020). SCO-GEM features an accuracy of 96.5% and 74.5% (Figure 1C) in predicting correct phenotypes for growth environments and knockout mutants, respectively, yielding in total a Matthews coefficient of correlation of 0.53 with the test data previously described (Kumelj et al., 2019).

With the recently published iKS1317 model (Kumelj et al., 2019) as a starting point, SCO-GEM was first developed by including genes, reactions, and metabolites from the equally recently published models iAA1259 (Amara et al., 2018) and SCO4 (Wang et al., 2018). The curations from iAA1259 were primarily related to coelomicin P1, butyrolactone, xylan, and cellulose pathways, whereas the 377 reactions added to SCO-GEM from SCO4 were scattered across a large range of different subsystems, covering both primary and secondary metabolism (Figure S1). Subsequent to merging the existing *S. coelicolor* GEMs, we performed a number of further curations of the model (Figure 1A): including improvement of annotations, both in terms of coverage and number of different databases, e.g., KEGG (Kanehisa, 2000; Kanehisa et al., 2019), BioCyC (Karp et al., 2019), ChEBI (Hastings et al., 2016), and MetaNetX (Moretti et al., 2016). All reactions and metabolites have been given identifiers according to the BiGG namespace (King et al., 2016), and all reactions are categorized into 15 different subsystems, covering 128 different pathways.

The biomass composition was curated to reflect estimated levels of prosthetic groups that are associated to cellular proteins. Proteomics data, as discussed later, were used to estimate protein levels, while UniProt (The UniProt Consortium, 2019) provided annotations of proteins with prosthetic groups, which was used to estimate overall prosthetic group levels (Data S1, Table 2).

Reaction Reversibility Updated for Almost a Third of Queried Reactions

The determination of reaction directionality and reversibility is an important step in a GEM reconstruction (Thiele and Palsson, 2010). However, the thermodynamic consistency of reactions was not considered in previous *S. coelicolor* models. We calculated Gibbs free energy changes for 770 of the 2,612 model reactions (Data S1, Table 3) using eQuilibrator (Fiamholz et al., 2012) and found hardly any consistency between the calculated change in Gibbs free energy and the reversibility previously assigned to the model reactions (Figure 1D). To address this issue we decided to reassign the reversibility of the model reactions by using a relatively lenient threshold of -30 kJ/mol to classify a reaction as irreversible (Bar-Even et al., 2012; Feist et al., 2007), with the intent not to over-constrain the model (Figure 1E). The proposed changes in reversibility were evaluated against growth and knockout data (Kumelj et al., 2019), discarding 61 of the 332 proposed reactions, and consequentially, the flux bounds of 271 reactions were modified (see Transparent Methods). In addition, all ATP-driven reactions were manually curated and generally assumed irreversible unless they had an estimated positive change in Gibbs free energy or were known to be reversible. Examples of this include nucleoside diphosphate kinase (Chakrabarty, 1998) and ATP synthase (Yoshida et al., 2001). The manual curation of ATP-driven reactions led to a change in reversibility for 56 reactions.

Curation of Transport Reactions

As transport reactions have previously not been extensively curated in *S. coelicolor* models, we performed a thorough curation of transporters by querying various databases and BLAST analysis as detailed in Methods. This culminated in adding 43 new transport reactions and updating 39 of the 262 existing reactions in SCO-GEM (Figure 1F; Data S1, Table 4). The majority of the transporters comprise primary active transport proteins and secondary carriers (46%), in accordance with previous work (Getsin et al., 2013). Most primary active transporters are ATP-binding cassette (ABC) transporters (30%), whereas proton symports (30%) dominate the secondary carriers.

Development of the Enzyme-Constrained Model EcSCO-GEM

To include explicit constraints regarding enzymes catalyzing metabolic reactions, the GECKO formalism (Sanchez et al., 2017) was applied to consider that catalyzing capacity is constrained by enzyme turnover rates (k_{cat}) and abundances. The GECKO toolbox modifies the structure of an existing GEM to integrate turnover rates and proteome data. Consequentially, this constrains the range of estimated fluxes to a biologically feasible range as determined by the amount and efficiency of each enzyme. Note that this approach regards the maximum catalytic activities but does not consider other kinetic parameters such as affinity constants. The overall flux variability of the resulting enzyme-constrained model (EcSCO-GEM) is drastically reduced compared with the classic genome-scale model (Figure 1G), particularly due to the

considerably reduced fraction of reactions that have very high (10^1) flux variability. As reactions with high variability result in low certainty in the estimated fluxes, the observed reduction in flux variability is therefore a qualitative measure of the increased accuracy achieved by constraining the range of possible fluxes to those satisfying the limitation in protein allocation.

In our endeavor to describe the metabolic differences between M145 and M1152 we generated in total 17 time- and strain-specific enzyme-constrained models by combining EcSco-GEM with estimated growth, secretion, and uptake rates, as well as proteome data from cultivations that are detailed and analyzed later in the article.

Framework for Further Development of Sco-GEM by the Community

The Sco-GEM model is hosted as an open repository as suggested by memote, a recently developed tool for transparent and collaborative model development (Lieven et al., 2020). The memote tool is incorporated in the repository through Travis CI and tracks the model development on every change of the model. Sco-GEM v1.2.0 achieved a memote score of 77%, which is superior to that achieved by any previous model of *S. coelicolor* (Figure 1B; Supplemental Information).

Hosting Sco-GEM on GitHub with memote integration ensures continuous quality control and enables public insight into all aspects of model reconstruction and curation: any user can report errors or suggest changes through issues and pull requests. As contributions to the model development are fully trackable and can therefore be credited fairly, Sco-GEM is positioned as a community model that we envision to be continuously updated and widely used by the *S. coelicolor* research community. Although the major steps of model reconstruction have been detailed in the preceding sections, every detail of the process and every iteration of the model is accessible on the public model repository at <https://github.com/SysBioChalmers/Sco-GEM>.

In the remaining parts of the Results section, we have applied Sco-GEM along with transcriptome and proteome data, to study and compare the responses of *S. coelicolor* M145 and M1152 to phosphate depletion on a systems level and for the first time provide detailed insight into the distinct physiological features of engineered “superhost” strain M1152, which will be of value for its further development.

Random Sampling of Enzyme-Constrained GEMs Capture Metabolic Rearrangements in Response to Phosphate Depletion in M145

To evaluate whether the (Ec)Sco-GEM models can simulate behaviors of *S. coelicolor* metabolism, we analyzed time course sampled cultivations of secondary metabolite-producing strain M145 using the generated models. For this purpose, *S. coelicolor* M145 was cultivated in batch fermentations using standardized protocols reported earlier (Wentzel et al., 2012a). Cultures were sampled for “omics” data, as well as substrate utilization and secondary metabolite measurements to identify regulatory, proteomic, and metabolic changes during the metabolic switch. The online and offline measurements showed that phosphate depletion in the cultivation medium was reached approximately 35 h after inoculation. Shortly after, the culture growth ceased, and first Red and subsequently Act were detected in the culture medium (Figures 2A and 2B). Act levels were determined by measuring the amount of total blue pigments because this covers both the intracellular and secreted variants of actinorhodin, and is considered to be the preferred method (Bystrykh et al., 1996; Wentzel et al., 2012a). Both D-glucose and L-glutamate were consumed concomitantly, and their consumption continued after phosphate depletion, whereas both remained in excess until the end of cultivation. Note that *Streptomyces* can utilize intracellular phosphate storages after the medium is phosphate depleted (Smirnov et al., 2015). The RNA sequencing (RNA-seq) and untargeted proteomic data were analyzed in the light of previous studies (Nieselt et al., 2010; Thomas et al., 2012) and were in good agreement with data previously obtained from microarrays or targeted proteomics (Alam et al., 2010; Nieselt et al., 2010) (Figures 2C and S2). This confirmed the high reproducibility of the experiments across independent cultivations and high reliability of the chosen cultivation and analytic procedures (Figure 2).

The proteome data and calculated uptake/secretion rates (Table S1) were incorporated into EcSco-GEM to yield time-specific metabolic models of M145, giving insight on the changes occurring in the metabolic activity of different pathways during batch cultivation. Metabolic fluxes were estimated using an unbiased approach of random sampling, as alternative to optimization of a well-defined cellular objective used in flux balance analysis (Orth et al., 2010). It is possible that *S. coelicolor* is wired to maximize its growth

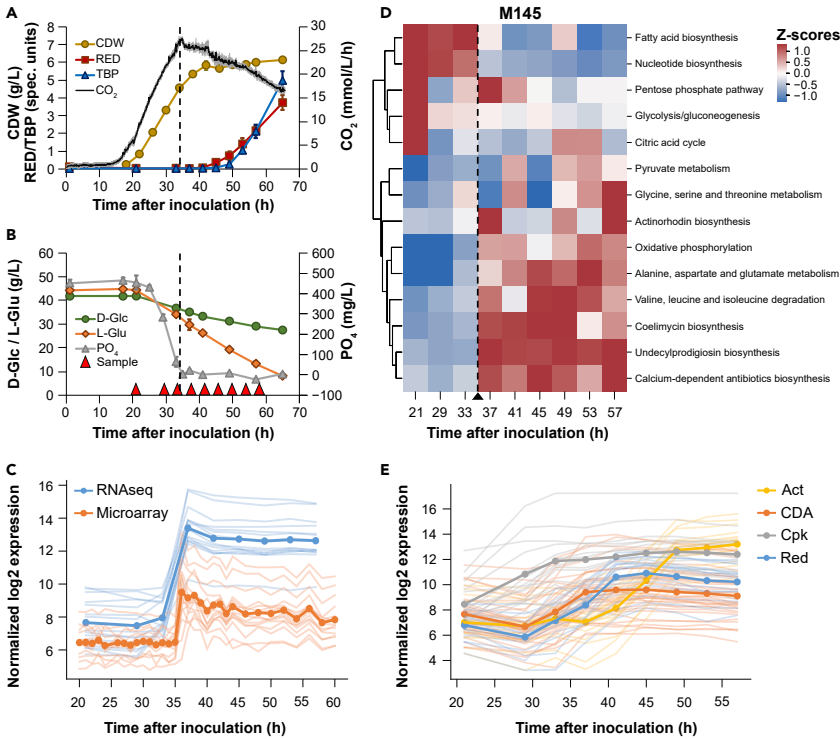


Figure 2. Batch Cultivation of *S. Coelicolor* M145 and the Effect of Phosphate Depletion

(A and B) Compounds produced (A) and consumed (B) during batch fermentation of *S. coelicolor* M145. Time points for sampling for transcriptome and proteome analysis are indicated with red triangles. The dashed vertical line indicates when phosphate in the medium has been depleted. Error bars are standard deviations of three biological replicates. CDW, cell dry weight; Red, undecylprodigiosin; TBP, total blue pigments/actinorhodins; CO₂, volume-corrected respiration; D-Glc, D-glucose; L-Glu, L-glutamate; PO₄, phosphate.

(C) Comparison of previously published microarray data (Nieselt et al., 2010) and RNA-seq data (this study) for genes previously found to respond to phosphate depletion (Nieselt et al., 2010). The transparent lines correspond to individual genes, whereas the bold lines represent the average expression level for each dataset.

(D) Clustered heatmap of CO₂-normalized Z-scores for each of the top 10 varying pathways plus the pathways for the four major BGCs in M145, as revealed by simulations with the proteomics-integrated EcSco-GEM model. The pathways are sorted based on hierarchical clustering to facilitate visual interpretation of similarity between pathways. The dashed vertical line indicates the time point of the metabolic switch.

(E) RNA-seq data of the four major BGCs show the onset of biosynthesis of actinorhodin (Act), calcium-dependent antibiotic (CDA), coelimycin P1 (Cpk), and undecylprodigiosin (Red) at different time points during the batch fermentations of M145.

rate before phosphate depletion, but after the metabolic switch, it is difficult to define a clear cellular objective. We applied an approach that samples the vertices of the solution space (Bordel et al., 2010) and used their mean values to compare the metabolic fluxes between the two strains and between different time points. The variation in predicted fluxes through different pathways in M145 is an initial validation of the approach (Figure 2D): the most drastic change in fluxes occur in response to phosphate depletion, in agreement with observations in the transcriptome, metabolome, and proteome (Nieselt et al., 2010; Thomas et al., 2012; Wentzel et al., 2012b).

The response to phosphate depletion from the medium is achieved by a set of genes, positively regulated by PhoP, that are involved in phosphate scavenging, uptake, and saving (Martin et al., 2012; Martin-Martin

et al., 2018; Sola-Landa et al., 2003). In our cultivations the metabolic switch can be readily identified from the RNA-seq data by the rapid upregulation of this regulon after 35 h of cultivation in M145 (Figure 2C), thereby corroborating the model simulations (Figure 2D) and providing a more detailed picture of the underlying regulation. PhoP also represses nitrogen assimilation (Martin et al., 2017), which can partly explain the change in amino acids metabolism after phosphate depletion (Figure 2D). Indeed, from the RNA-seq data we find that glutamate import, the glutamate sensing system *gluR-gluK* (Li et al., 2017), *glnR* (Fink et al., 2002), and *glnA* are downregulated immediately subsequent to phosphate depletion (Figure S3). As PhoP is also known to regulate negatively the biosynthesis of secondary metabolites, the switching of its expression likely delays these pathways (Martin, 2004; Martin et al., 2017). However, after 37 h of cultivation the upregulation of the *cda* and *red* genes was observed, whereas that of the *act* genes was initiated at 41 h (Figure 2E). Production of Red and Act was measurable in the culture medium after 41 and 49 h of cultivation, respectively (Figure 2A). The enzyme-constrained models predict an immediate increase in fluxes through the biosynthetic pathways for the four main compounds Act, Red, CDA, and coelomicyn P1 after the metabolic switch (Figure 2D).

The Onset of Secondary Metabolism Is Strongly Correlated with an Increase in Oxidative Phosphorylation and a Decrease in Fatty Acid Biosynthesis in M145

The metabolic switch was shown to be correlated with an enhanced degradation of branched-chain amino acids (valine, leucine, and isoleucine), an increase in oxidative phosphorylation, and a decrease in fatty acid biosynthesis (Figures 2D and S4). An active oxidative phosphorylation relies on an active tricarboxylic acid (TCA) cycle that generates reduced co-factors whose re-oxidation by the respiratory chain generates a proton gradient that drives ATP synthesis by the ATP synthase. The feeding of the TCA cycle requires acetyl-CoA, as well as nitrogen. Nitrogen likely originates from degradation of glutamate and branched-chain amino acids, whereas acetyl-CoA likely originates from glycolysis, as well as from the degradation of these amino acids as previously demonstrated (Stirrett et al., 2009). Indeed, the model predicts an increased flux through citrate synthase feeding acetyl-CoA into the TCA cycle (Figure S5A). The predicted increase in oxidative phosphorylation is supported by the RNA-seq data showing upregulation of enzymes belonging to the respiratory chain (Figure S5B). This is consistent with the clear correlation previously reported between high ATP/ADP ratio, resulting from an active oxidative phosphorylation, and actinorhodin production (Esnault et al., 2017). Furthermore, the consumption of acetyl-CoA by the TCA cycle to support the oxidative metabolism logically impairs fatty acids biosynthesis (Esnault et al., 2017).

The pentose phosphate pathway provides the main redox cofactor NADPH for polyketide biosynthesis, as well as to combat oxidative stress, and its model-predicted flux increase upon initiation of polyketide synthesis (Figure 2D) is in agreement with previous studies (Borodina et al., 2008; Jonsbu et al., 2001). A clear positive correlation was also noticed between the biosynthesis of alanine, aspartate, and glutamate, which are precursors for CDA and/or coelomicyn P1 (Figure 2D), and the biosynthesis of these antibiotics. Similar observations were made in the antibiotic-producing *Amycolatopsis* sp. (Gallo et al., 2010). Our EcSco-GEM model proved to be in good agreement with previously reported findings, indicating that it is able to capture *S. coelicolor* metabolic behavior.

Model-Assisted Characterization of Engineered *S. Coelicolor* M1152 and Its Responses to Phosphate Depletion

As detailed earlier, EcSco-GEM shed a new light on the metabolic switch in secondary metabolite-producing strain M145. *S. coelicolor* M1152 (Gomez-Escribano and Bibb, 2011) is an M145 derivative devoid of the four major BGCs and bearing a point mutation in the *rpoB* gene. A better systemic understanding of M1152 metabolism would benefit to its further development as a performing host. To do so, a comparative analysis of gene expression levels and metabolic fluxes was carried out in the strains M145 and M1152.

Batch cultivations of M1152 were performed using identical conditions and comparable sampling regimes as for M145 reported earlier. This enabled a direct comparison of the two strains at a systems level, revealing both expected and unexpected effects of the strains' genetic differences (Figure 3). As anticipated, the products of the Cpk, CDA, Red, and Act biosynthetic pathways were undetectable in M1152 (Figure 3A). As previously observed (Gomez-Escribano and Bibb, 2011), the growth rate of M1152 is reduced compared with M145 (0.15 h^{-1} versus 0.21 h^{-1} in the initial exponential growth phase), delaying phosphate depletion by M1152 to 47 h after inoculation (Figure 3B), 12 h after M145 (Figure 2B).

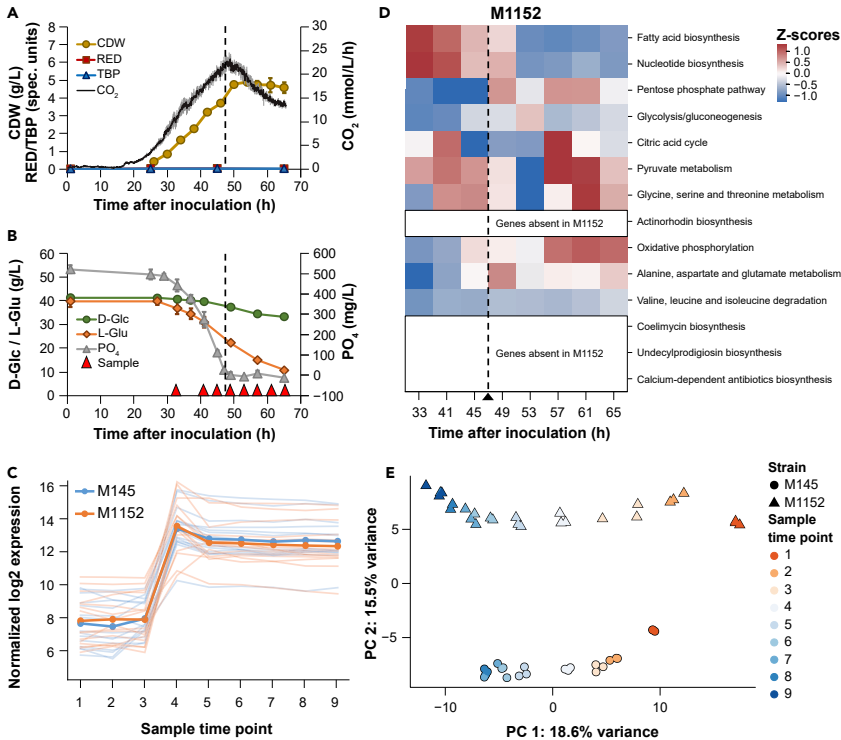


Figure 3. Batch Cultivation of *S. Coelicolor* M1152

(A and B) Compounds produced (A) and consumed (B) during batch fermentation of *S. coelicolor* M1152. Time points for sampling for transcriptome and proteome analysis are indicated with red triangles. The dashed vertical line indicates when phosphate in the medium has been depleted. Error bars are standard deviations of three biological replicates. CDW, cell dry weight; Red, undecylprodigiosin; TBP, total blue pigments/actinorhodins; CO₂, volume-corrected respiration; D-Glc, D-glucose; L-Glu, L-glutamate; PO₄, phosphate.

(C) Alignment of sample time points of M145 and M1152 cultivations based on the expression profiles of genes that were earlier found to respond to phosphate depletion with respect to the metabolic switch (Nieselt et al., 2010).

(D) Principle-component analysis of the proteomics data for M145 (triangles) and M1152 (circles), for each time point and culture. The first principal component separates the time points, whereas the second principal component separates the two strains.

(E) CO₂-normalized Z scores of pathway fluxes predicted by EcSco-GEM for 10 of the most varying pathways in M145 and M1152. To make this heatmap comparable to the results for M145 (Figure 2D), the data are standardized for both strains simultaneously and the row order is identical.

The sampling time points for proteome and transcriptome were adjusted accordingly (Figure 3B), enabling pairwise comparison of measurements between the two strains. Genes responsive to phosphate depletion, members of the PhoP regulon (Nieselt et al., 2010), were used to align the different sample datasets for M145 or M1152 (Figure 3C). Principle-component analysis of the proteome data confirms high consistency between corresponding biological replicates and incremental changes between sample points for both M145 and M1152 (mainly explained by principal component 1 (PC1): 18.6% variance, Figure 3E). A clear strain-dependent clustering of the data (PC2: 15.5% variance) indicates globally significant differences at the protein level. EcSco-GEM was subsequently used to create time-specific metabolic models from proteome data and estimated rates (Table S2) and predict metabolic changes in M1152. Interestingly we find that most patterns in M145 are retained in M1152 (Figure 3D): fatty acid and nucleotide biosynthesis is still downregulated after phosphate depletion, and similar trends of upregulation at later time points are observed for oxidative phosphorylation, glycine, serine and threonine, and pyruvate metabolism. It is

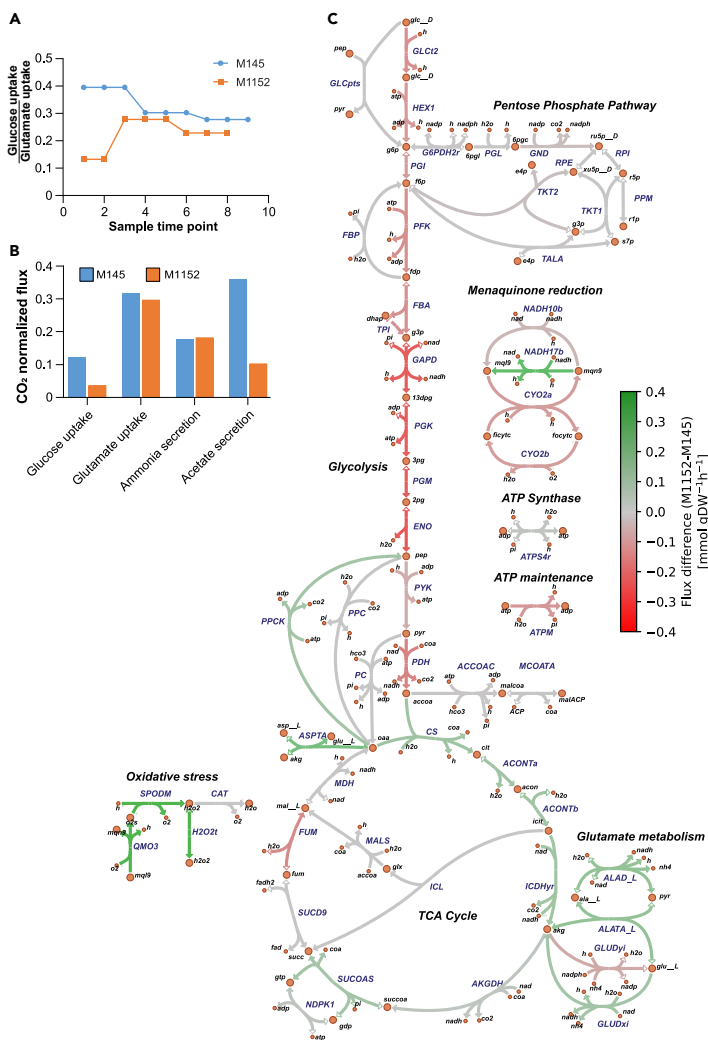


Figure 4. Predicted Carbon Fluxes in M145 and M1152

(A) The ratio between estimated uptake rates of glucose and glutamate for each sample time point for M145 and M1152 shows that M1152 acquires a smaller part of its carbon from glucose compared with M145.

(B) Bar chart showing CO₂-normalized fluxes for the second sampling time point for M145 and M1152, i.e., after 29 and 41 h, respectively. There is a clear difference in the uptake of glucose and production of acetate, whereas the rates are comparable for the consumption of glutamate and secretion of ammonium.

(C) Comparison of predicted fluxes for the second sampling time points shows clear differences between the two strains in their relative utilization of the glycolysis and TCA cycle. The strength of the color of the lines corresponds to the flux difference between the strains; green reactions have higher flux in M1152, and red reactions have higher flux in M145.

striking that the upregulation of the branched-chain amino acid degradation and the alanine, aspartate, and glutamate metabolism seen as a response to phosphate depletion in M145 are absent in M1152.

The different glutamate and glucose consumption rates of M145 and M1152 (Figures 4A and 4B) resulted in substantial metabolic differences between the two strains before phosphate depletion. During cultivation

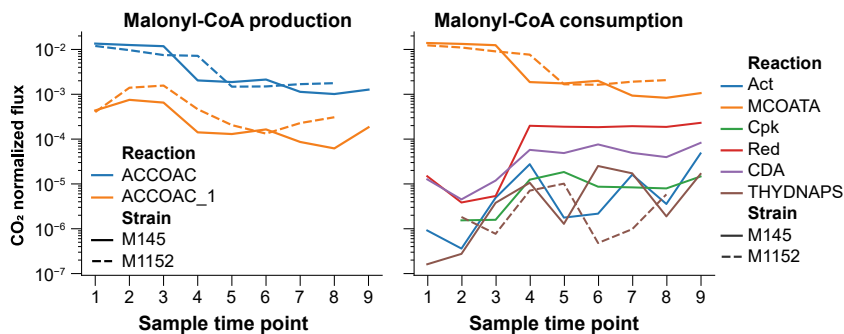


Figure 5. Production and Consumption of Malonyl-CoA as the Branching Point between Fatty Acid Biosynthesis and Production of Polyketides

Both panels display CO₂-normalized fluxes for both M145 and M1152 for all sampling time points as predicted by EcSco-GEM. The left panel shows the sources of malonyl-CoA, namely, acetyl-CoA carboxylase (ACCOAT; blue) and acetyl-CoA carboxyltransferase (ACCOAT_1; orange). We observe a downregulation of the malonyl-CoA production after the metabolic switch (between time points 3 and 4) in both strains. The right panel presents reactions consuming malonyl-CoA. The consumption is dominated by malonyl-CoA-ACP transacylase (MCOATA) leading to biosynthesis of fatty acids. The other drains for malonyl-CoA are the pathways encoded by the four major BGCs (Act, Cpk, Red, and CDA) in addition to biflavinol synthase (THYDNAPS).

on SSBM-P medium, where glutamate is the sole nitrogen source, glucose and glutamate are co-consumed. M1152, as M1146 (Esnault et al., 2017), has an increased growth yield on glucose compared with M145 (Figure S6). It thus obtains a larger share of its carbon from glutamate (Figures 4A and 4B) and has consequently also a higher nitrogen availability than M145. The increased nitrogen availability does, however, not increase the secretion of secondary metabolites. A reduced flux through glycolysis has also been reported previously for strain M1146 (Coze et al., 2013). This might be an effect of the predicted increased concentration of ATP in M1146 compared with M145, which inhibits glucose uptake and phosphofructokinase (Coze et al., 2013; Esnault et al., 2017). As Act was proposed to act as an electron acceptor reducing the efficiency of the oxidative phosphorylation, it is suggested that the lack of Act in M1146 causes the elevated ATP levels (Esnault et al., 2017). However, we find the largest difference in glycolytic flux at early time points, before phosphate depletion and Act production in M145, proving that Act itself cannot explain this observation.

The EcSco-GEM predicts the consequences of the reduced glucose uptake of M1152 on its central carbon metabolism, as displayed by mapping relative reaction fluxes from the second sampling time point onto a map of the central carbon metabolism in *Streptomyces* (Figure 4C). The map is based on the reaction network in Sco-GEM and created using Escher (King et al., 2015). A less-active glycolysis in M1152 than in M145 leads to a lower carbon flow toward acetyl-CoA and thus lower excretion of acetate compared with M145 (Figure 4B). Furthermore, EcSco-GEM reveals an increased flux from glutamate to alpha-ketoglutarate. Indeed, a fraction of the pool of oxaloacetate might be converted into alpha-ketoglutarate by aspartate transaminase to feed the TCA cycle. The rest might be converted into phosphoenolpyruvate (PEP) by PEP carboxykinase for gluconeogenesis because PEP carboxykinase was shown to carry higher fluxes in M1152 than in M145 (Figure 4C).

As recent studies have demonstrated a negative correlation and a competition for common precursors between secondary metabolite and triacylglycerol (TAG) biosynthesis in *S. lividans* and *S. coelicolor* (Crane et al., 2012; Esnault et al., 2017; Millan-Oropeza et al., 2017), one can speculate that the acetyl-CoA/malonyl-CoA units yielded by glycolysis for the biosynthesis of antibiotics in M145 are being used for enhanced growth and/or fatty acids and TAG biosynthesis in M1152. However, this is likely not the case, as M1152 has rather a reduced growth rate compared with M145, and fatty acid biosynthesis remains down-regulated after the switch (Figure 5). Malonyl-CoA is predominantly shuttled toward fatty acid biosynthesis through malonyl-CoA-ACP transacylase, and this consumption seems to be well balanced by the amount of malonyl-CoA produced by acetyl-CoA carboxylase. It is noteworthy that the flux toward this acetyl-CoA/

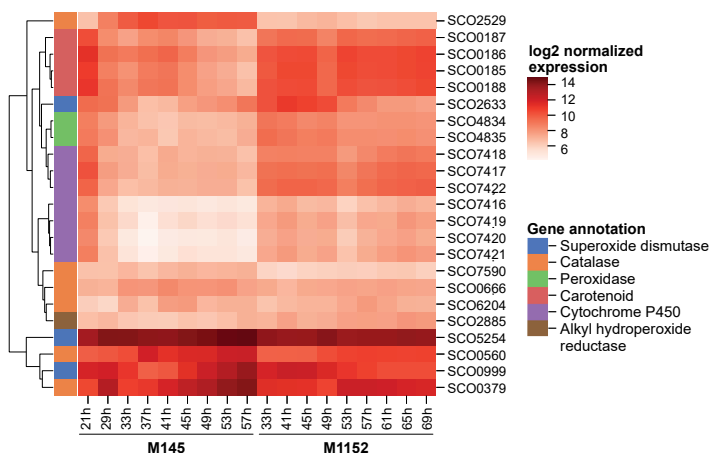


Figure 6. Heatmap Displaying Log-Transformed RNA-Seq Data of Genes Associated with Oxidative Stress

The genes included are related to oxidative stress and either present in Sco-GEM or within the 499 differentially expressed genes. These genes are categorized based on their functional annotation to distinguish differences and similarities between these functional groups. To further enhance visual interpretation the genes are ordered based on hierarchical clustering to align genes with similar expression profiles across M145 and M1152 next to each other.

malonyl-CoA drain is 3- to 6-fold larger than the total flux going into secondary metabolite biosynthesis, even after the metabolic switch. We thus propose that together with enhanced nitrogen availability, acetyl-CoA made available from the deletion of these BGCs is used to feed the TCA cycle to support the oxidative metabolism in M1152. This would generate oxidative stress whose toxic effects might be responsible for the growth delay of this strain.

Transcriptome Analysis Reveals Differential Expression of Global Regulators

Although the proteome data are an integral part of the EcSco-GEM models, RNA-seq data were used to both verify the trends and to gain further insights into the regulatory changes that are not captured by the metabolic models. As the proteomic data, the RNA-seq data showed large global differences between M1152 and M145, revealing 499 differentially expressed genes with a significance threshold of $p < 0.01$.

Unsupervised clustering of the significantly changed genes reveal differences in regulatory systems related to redox regulation, signaling, and secondary metabolism. The significantly changed genes were clustered into seven groups with K-means clustering, with clusters 1–3 containing genes that are upregulated in M1152 compared with M145, and clusters 4–7 vice versa (Figure S7A; Data S2). A Gene Ontology (Ashburner et al., 2000; The Gene Ontology Consortium, 2019) enrichment analysis of the seven clusters was conducted to identify upregulated processes in each of the two strains (Figure S8, cf. Figure S7A).

The enriched processes upregulated in M1152 point to increased oxidative stress (Figure S8): antioxidant and peroxidase activity (SCO2633 [sodF]; SCO4834–35) in addition to biosynthesis of carotenoid (SCO0185–SCO0188), a known antioxidant (Latifi et al., 2009; Stahl and Sies, 2003). The putative proteins within the cytochrome-P450 family (SCO7416–SCO7422) found in cluster 1 might be linked not only to increased oxidative stress (Zangar et al., 2004) but also to oxidation of precursors used for the synthesis of macrolides (Lamb et al., 2003). Indeed, by comparing the time series expression levels for genes related to oxidative stress we observe that the majority of genes related to oxidative stress are upregulated in M1152 (Figure 6). These changes correlate to a more active oxidative metabolism, TCA cycle, and oxidative stress as predicted by Ec-ScoGEM (Figure 4).

In cluster 2 we find *scbA* (SCO6266) and its downstream gene *scbC* (SCO6267), which stands out by being almost 6-fold upregulated in M1152. This high expression level is likely due to the deletion of *scbR2* (SCO6286), the last gene selected to be part of the *cpk* BGC (Bednarz et al., 2019). Besides regulation of

the *cpk* cluster, ScbR2 binds upstream of several global regulators of development and secondary metabolism, including AfsK, SigR, NagE2, AtrA, AdpA, and ArgR (Li et al., 2015). It also acts together with ScbR to regulate ScbA, which produces the γ -butyrolactone SCB1. However, when looking at the genes regulated by ScbR (Li et al., 2015), we only observe a clear difference in expression for genes regulated by AfsR (phosphorylated by AfsK) (Horinouchi, 2003; Lee et al., 2002), whereas this is not the case for genes regulated by ArgR, AdpA, or ScbR itself (Figures S5C-S5F).

Among the genes upregulated in M145, in cluster 4 we find genes related to the redox-regulated transcription factor SoxR (Naseer et al., 2014), and a similar pattern is observed for the entire SoxR regulon (Figure S7B). SoxR is known to react directly to the presence of actinorhodin (Dela Cruz et al., 2010; Shin et al., 2011), and indeed, in M145 this group of genes follows the production profile of actinorhodin, whereas their expression remains low in M1152 as Act is not produced. The benzoquinone Act, as electron acceptor, is thought to reduce respiration efficiency and thus energy charge, as well as to combat oxidative stress (Esnault et al., 2017). Consistently, the RNA-seq data revealed that the ATP-synthase gene cluster (SCO5366–SCO5374) was upregulated almost 2-fold in M1152 compared with M145, most prominently in the stationary phase during Act production (Figure S7C). This agrees with observations in the M1146 strain (Coze et al., 2013). Cluster 4 also contains the genes directly up- and downstream of the deleted actinorhodin BGC in M1152 (SCO5071–SCO5072, encoding 3-hydroxyacyl-CoA dehydrogenase, and SCO5091–SCO5092, encoding a two-component flavin-dependent monooxygenase system) (Valton et al., 2008). In clusters 5, 6, and 7 we find genes with reduced expression in M1152, and the enriched processes are related to cellular and iron ion homeostasis, development, signaling, and morphology. This corresponds to the delayed sporulation observed for M1152 (Gomez-Escribano and Bibb, 2011).

Elevated Expression of Ribosomal Proteins in M1152 after Phosphate Depletion

An increased transcription of genes encoding ribosomal proteins could be observed in M1152 after phosphate depletion (Figure S7D). The *rpoB* mutation of the RNA polymerase present in M1152 is thought to induce a conformational change mimicking the binding of guanosine tetraphosphate (ppGpp) to this enzyme (Hu et al., 2002). ppGpp is synthesized in response to nutritional stress and reduces the transcription of genes related to active growth, such as genes encoding ribosomal RNAs and ribosomal proteins (Burgos et al., 2017), whereas it upregulates those involved in development/differentiation and antibiotic production (Hesketh et al., 2007; Srivatsan and Wang, 2008). In consequence the upregulation of ribosomal proteins was unexpected in M1152, especially because the expression of the ppGpp regulon was not found to be significantly changed in M1152 (Figure S5G and S5H). We hypothesize that the ribosomal upregulation originates from the higher ATP content of M1152 compared with M145 post phosphate depletion, as high nucleoside triphosphate levels are known to have a positive impact on ribosome synthesis (Gaal et al., 1997). Such difference in ribosomal protein expression is mainly seen in the antibiotic production phase and correlated with production of Act in M145, which has a negative impact on the energetic state of the cell (Esnault et al., 2017).

Reduced Production of the Polyketide Germicidin in M1152

One could reasonably anticipate that the production of a secondary metabolite would increase if other drains competing for same precursor compounds were removed from the organism by gene deletion. However, the production rate of the polyketides germicidin A and B (Chemler et al., 2012), autologous to both M145 and M1152, were reduced in M1152 by 92% and 82% for germicidin A and B, respectively (Figure 7). This could be explained by the more active oxidative metabolism of M1152 compared with M145, as suggested by the enzyme-constrained model (Figure 4) and supported by the upregulation of genes associated with oxidative stress (Figure 6). In M1152 the pool of acetyl-CoA rather feeds the TCA cycle instead of being directed toward germicidin biosynthesis.

To further elucidate the cause of the reduced production in M1152, we also measured germicidin production in the intermediate strain M1146 (Figures 7 and S7E), which does not feature the *rpoB* mutation but is missing the four BGCs also deleted in M1152 (Gomez-Escribano and Bibb, 2011). The production rate of germicidin A and B in M1146 was found to be reduced by 27% and 25%, respectively, compared with M145. When compared with the strong reduction in germicidin production that can be assigned to the *rpoB* mutation in M1152, removal of only the four BGCs in M1146 has a moderate effect on germicidin production. This conforms with the minor contribution of the BGCs compared with fatty acid biosynthesis on the total consumption of malonyl-CoA (Figure 5). Nonetheless, it remains contradictory that the removal of polyketide precursor drains negatively impacts the production of other polyketides.

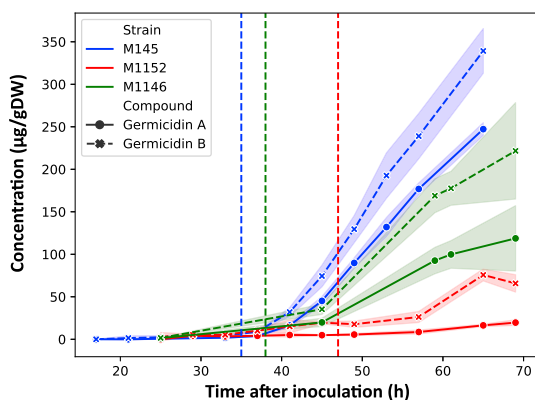


Figure 7. Concentrations of Germicidin A and B Produced by M145, M1146, and M1152

The concentrations are normalized by the biomass of each strain. The shaded regions display the uncertainty range (± 1 standard deviation) based on three replicate cultivations. Note that the growth rate is different between the strains, displayed by the vertical lines representing phosphate depletion at 35, 38, and 47 h for M145, M1146, and M1152, respectively.

DISCUSSION

In this work, we carried out a multi-omics study to compare the metabolic changes of *Streptomyces coelicolor* M145 and the BGC deletion mutant M1152 during batch fermentation. The defined cultivation medium used in this work was chosen because it supports sufficient growth and a delayed, well-defined onset of secondary metabolism, necessary to study the metabolic switch (Wentzel et al., 2012a). We aimed at defining the metabolic features differing between the two strains, both during exponential growth and stationary phase after phosphate depletion.

To achieve this from a systems biology perspective, we combined time course sampled cultivation and transcriptome analysis with enzyme-constrained genome-scale models generated with proteome data. Such genome-scale models are extensively used to connect transcriptome and proteome data to metabolic fluxes. Leveraging metabolic simulations to contextualize transcriptional changes is mainly impacted by the quality of the computational model used. Here, two teams joined efforts to improve a consensus model of *S. coelicolor*, yielding a comprehensive model useful for the scientific community.

Genome-Scale Models Provide Hypothesis for Slow Growth of M1152

The reduced growth rate of M1152 is correlated with reduced glucose uptake and enhanced glutamate uptake compared with M145. This is expected to lead to a less active glycolysis but a more active TCA cycle, and thus, a more active oxidative metabolism in M1152 compared with M145. An active oxidative metabolism is known to generate oxidative stress, and indeed, the *in vivo* data, as well as the genome-scale model, predict an increased oxidative stress in M1152. The toxicity of oxidative stress might, at least in part, be responsible for the growth delay of M1152, whereas the *rpoB* mutation may add to this phenotype, because one of the functions of the ppGpp-associated RNA polymerase is to promote a growth arrest in conditions of nutritional stress.

Further Development May Improve M1152 as Host for Heterologous Expression

The strain M1152 has several advantages as a host for heterologous production of secondary metabolites. The deletion of the four major BGCs not only removes presumed competing drains for valuable precursors but also generates a clean background to ease the identification of novel products by mass spectrometry. M1152 has already been proved to be more efficient than M145 and M1146 in heterologous production of the nitrogen-containing antibiotics chloramphenicol and congocidine, as well as Act production from reintroduction of its BGC (Gomez-Escribano and Bibb, 2011). Strains M1146 and M1152 produce, respectively, 3- to 5-fold and 20- to 40-fold more chloramphenicol and congocidine from respective heterologous clusters than M145, a clear demonstration of the huge impact on production due to the *rpoB* mutation. Although this contrasts with our data

showing that M1152 has the lowest production of germicidin, it is relevant to note that chloramphenicol and congocidine are non-ribosomal peptide synthases relying on amino acids rather than malonyl-CoA as precursors. Although our data show reduced degradation of branched-chain amino acids and metabolism of alanine, aspartate, and glutamate as the clearest metabolic divergence upon phosphate depletion in M1152, as congocidine and chloramphenicol are based on aromatic amino acids the connection to increased production of these NRPs is not obvious. Another option is that the increased oxidative metabolism in M1152 provides more redox cofactors to drive the synthesis of these molecules. If competition for valuable precursors was rate limiting, the absence of the polyketides actinorhodin and coelimycin P1 should at least enhance the production of germicidin, all being dependent on malonyl-CoA. Moreover, differences in cultivation media further convolute cross-study comparisons: the aforementioned study use a complex growth medium, whereas we used a defined medium with glucose and glutamate, which has previously been optimized for studying the metabolic switch (Wentzel et al., 2012a).

Furthermore, (re-)introduction of a (secondary) copy of germicidin synthase gene *gcs* in strains M1152 and M1317—derived from M1152 by additional removal of three type III PKS genes including *gcs*—gave a 7.8- and 10.7-fold increase in germicidin production, respectively, compared with M1152 with only the native copy of *gcs* (Thanapipatsiri et al., 2015). Thus, the largest increase in production is not achieved by removal of competing precursor drains, but rather effected by the re-introduction of *gcs*, probably because expression of the inserted gene is not constrained by the same regulatory mechanism as the native gene.

Although earlier work has suggested a competition for common precursors between fatty acids and secondary metabolites biosynthesis (Craney et al., 2012), our results suggest that other approaches than deletion of competing precursor drains may be more efficient in the development of an optimized expression host, and it seems likely that different classes of BGCs may require different hosts for maximal production. Our comparative analysis of M145 and M1152 supports this development, not only as a systemic description connecting non-trivial associations between phenotypic, genetic, and metabolic differences but also by highlighting cellular processes that seem to be out of balance in M1152. These include upregulation of ribosomal genes, most likely an effect of the *rpoB* mutation, and increased oxidative metabolism and oxidative stress. As Act itself works as an electron acceptor one may hypothesize that its presence could relieve some of this stress. Another approach is to reintroduce *scbR2* to avoid influencing the related regulators of development and secondary metabolism.

Although *S. coelicolor* seems to have a complex and not fully elucidated regulatory system, several studies have shown that manipulation of regulatory genes can affect the production of secondary metabolites (Jones et al., 2011; Kim et al., 2003; Okamoto et al., 2003; Rodriguez et al., 2012). The complex regulation of secondary metabolite biosynthesis makes rational strain design difficult (Liu et al., 2013), but black-box approaches including random mutations and screening are still viable approaches for strain development (van den Berg et al., 2008; Crook and Alper, 2012). The Sco-GEM can aid this development by predicting the impact of these genetic alterations and to interpret “omics” data.

Limitations of the Study

We have performed a thorough comparison, of *S. coelicolor* M1145 and M1152, but to fully attribute changes in metabolism to the different genetic modifications as well as to unravel possible epistatic interactions we believe that a comprehensive analysis that also includes the intermediate strain M1146, and possibly also an M145 strain featuring only the *rpoB* mutation (Xu et al., 2002), will be necessary.

Resource Availability

Lead Contact

Further information and requests for resources and reagents should be directed to and will be fulfilled by the Lead Contact, Eduard J Kerkhoven (eduardk@chalmers.se).

Materials Availability

This study did not generate new unique reagents.

Data and Code Availability

The models and scripts generated during this study are available at GitHub (<https://github.com/SysBioChalmers/Sco-GEM>). Here, the latest version of the Sco-GEM is available in both YAML and

SBML level 3 Version 1. In addition, users can contribute to further model development by posting issues or suggest changes. The proteomics data are available from ProteomeXchange: PXD013178 via the PRIDE partner repository (Perez-Riverol et al., 2019). Normalized proteome data are also available in Data S3. The transcriptomics data are available from NCBI GEO: GSE132487 (M145) and GSE132488 (M1152). Normalized counts are also found in Data S4.

METHODS

All methods can be found in the accompanying [Transparent Methods supplemental file](#).

SUPPLEMENTAL INFORMATION

Supplemental Information can be found online at <https://doi.org/10.1016/j.isci.2020.101525>.

ACKNOWLEDGMENTS

The authors would like to acknowledge Bogdan I. Florea of Leiden University, Leiden, Netherlands, for running and monitoring the proteome measurements and the bio-organic synthesis group at Leiden University for providing the opportunity to use their instrumentation. The authors would also like to acknowledge co-workers at SINTEF Industry, Trondheim, Norway: Ingemar Nærdal, Anna Lewin, and Kari Hjelen for running the batch fermentations and Anna Nordborg, Janne Beate Øiaas, and Tone Haugen for performing offline analyses and the germicidin analytics. The RNA-seq sequencing was carried out by c.ATG, Tübingen, Germany.

This study was conducted in the frame of ERA-net for Applied Systems Biology (ERA-SysAPP) project SYSTERACT and the project INBioPharm of the Center for Digital Live Norway (Research Council of Norway grant no. 248885), with additional support of SINTEF internal funding.

AUTHOR CONTRIBUTIONS

Conceptualization, E.J.K., E.A., A.W., S.S., A.S.-Y., and T.K.; Methodology and Software, E.J.K., S.S., A.S.-Y., and T.K.; Validation and Formal Analysis, C.D., K.N., D.v.D., S.S., T.K., and E.J.K. Investigation, T.K., A.W., S.S., E.J.K.; Data Curation, S.S. and T.K.; Writing – Original Draft, S.S., T.K., D.V.D., C.D., and K.N.; Writing – Review & Editing, all authors; Visualization, S.S., E.K., C.D., and D.v.D.; Supervision, A.W., E.A., E.J.K., and G.P.v.W.; Project Administration, A.W.; Funding Acquisition: A.W., E.J.K., E.A., and G.P.v.W.

DECLARATION OF INTERESTS

The authors declare no competing interests.

Received: May 8, 2020

Revised: July 19, 2020

Accepted: August 31, 2020

Published: September 25, 2020

REFERENCES

- Alam, M.T., Merlo, M.E., (stream), T.S.C., Hodgson, D.A., Wellington, E.M., Takano, E., and Breitling, R. (2010). Metabolic modeling and analysis of the metabolic switch in *Streptomyces coelicolor*. *BMC Genomics* 11, 202.
- Amara, A., Takano, E., and Breitling, R. (2018). Development and validation of an updated computational model of *Streptomyces coelicolor* primary and secondary metabolism. *BMC Genomics* 19, 519.
- Ashburner, M., Ball, C.A., Blake, J.A., Botstein, D., Butler, H., Cherry, J.M., Davis, A.P., Dolinski, K., Dwight, S.S., Eppig, J.T., et al. (2000). Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat. Genet.* 25, 25–29.
- Aung, H.W., Henry, S.A., and Walker, L.P. (2013). Revising the Representation of fatty acid, glycerolipid, and glycerophospholipid metabolism in the consensus model of yeast metabolism. *Ind. Biotechnol.* 9, 215–228.
- Bar-Even, A., Flamholz, A., Noor, E., and Milo, R. (2012). Thermodynamic constraints shape the structure of carbon fixation pathways. *Biochim. Biophys. Acta* 1817, 1646–1659.
- Battke, F., Symons, S., and Nieselt, K. (2010). Mayday–integrative analytics for expression data. *BMC Bioinformatics* 11, 121.
- Bednarz, B., Kotowska, M., and Pawlik, K.J. (2019). Multi-level regulation of coelimirin synthesis in *Streptomyces coelicolor* A3(2). *Appl. Microbiol. Biotechnol.* 103, 6423–6434.
- Bentley, S.D., Chater, K.F., Cerdeño-Tárraga, A.-M.A.-M., Challis, G.L., Thomson, N.R., James, K.D., Harris, D.E., Quail, M.A., Kieser, H., Harper, D., et al. (2002). Complete genome sequence of the model actinomycete *Streptomyces coelicolor* A3(2). *Nature* 417, 141–147.
- van den Berg, M.A., Albarg, R., Albermann, K., Badger, J.H., Daran, J.-M., Driessen, A.J.M., Garcia-Estrada, C., Fedorova, N.D., Harris, D.M., Heijne, W.H.M., et al. (2008). Genome sequencing and analysis of the filamentous fungus *Penicillium chrysogenum*. *Nat. Biotechnol.* 26, 1161–1168.

- Bordel, S., Agren, R., and Nielsen, J. (2010). Sampling the solution space in genome-scale metabolic networks reveals transcriptional regulation in key enzymes. *PLoS Comput. Biol.* 6, e1000859.
- Borodina, I., Krabben, P., and Nielsen, J. (2005). Genome-scale analysis of *Streptomyces coelicolor* A3(2) metabolism. *Genome Res.* 15, 820–829.
- Borodina, I., Siebring, J., Zhang, J., Smith, C.P., Keulen, G.van, Dijkhuizen, L., and Nielsen, J. (2008). Antibiotic overproduction in *Streptomyces coelicolor* A3(2) mediated by phosphofructokinase deletion. *J. Biol. Chem.* 283, 25186–25199.
- Braesel, J., Tran, T.A., and Eustáquio, A.S. (2019). Heterologous expression of the diazaquinomycin biosynthetic gene cluster. *J. Ind. Microbiol. Biotechnol.* 46, 1359–1364.
- Burgos, H.L., O'Connor, K., Sanchez-Vazquez, P., and Gourse, R.L. (2017). Roles of transcriptional and translational control mechanisms in regulation of ribosomal protein synthesis in *Escherichia coli*. *J. Bacteriol.* 199, e00407-17.
- Bystrykh, L.V., Fernández-Moreno, M.A., Herrema, J.K., Malpartida, F., Hopwood, D.A., and Dijkhuizen, L. (1996). Production of actinorhodin-related "blue pigments" by *Streptomyces coelicolor* A3(2). *J. Bacteriol.* 178, 2238–2244.
- Castro, J.F., Razmilic, V., Gomez-Escribano, J.P., Andrews, B., Asenjo, J.A., and Bibb, M.J. (2015). Identification and heterologous expression of the chaxamycin biosynthesis gene cluster from *Streptomyces leeuwenhoekii*. *Appl. Environ. Microbiol.* 81, 5820–5831.
- Chakrabarty, A.M. (1998). Nucleoside diphosphate kinase: role in bacterial growth, virulence, cell signalling and polysaccharide synthesis. *Mol. Microbiol.* 28, 875–882.
- Chandra, G., and Chater, K.F. (2014). Developmental biology of *Streptomyces* from the perspective of 100 actinobacterial genome sequences. *FEMS Microbiol. Rev.* 38, 345–379.
- Chemler, J.A., Buchholz, T.J., Geders, T.W., Akey, D.L., Rath, C.M., Chlipala, G.E., Smith, J.L., and Sherman, D.H. (2012). Biochemical and structural characterization of germicidin synthase: analysis of a type III polyketide synthase that employs acyl-ACP as a starter unit donor. *J. Am. Chem. Soc.* 134, 7359–7366.
- Coze, F., Gilard, F., Tcherkez, G., Virolle, M.-J., and Guyonvarch, A. (2013). Carbon-flux distribution within *Streptomyces coelicolor* metabolism: a comparison between the actinorhodin-producing strain M145 and its non-producing derivative M1146. *PLoS One* 8, e84151.
- Craney, A., Ozimok, C., Pimentel-Elardo, S.M., Capretta, A., and Nodwell, J.R. (2012). Chemical perturbation of secondary metabolism demonstrates important links to primary metabolism. *Chem. Biol.* 19, 1020–1027.
- Crook, N., and Alper, H. (2012). *Classical Strain Improvement. Engineering Complex Phenotypes in Industrial Strains* (John Wiley & Sons, Inc.). <http://dx.doi.org/10.1002/9781118433034.Ch1>.
- Dela Cruz, R., Gao, Y., Penumetcha, S., Sheplock, R., Weng, K., and Chander, M. (2010). Expression of the *Streptomyces coelicolor* SoxR regulon is intimately linked with actinorhodin production. *J. Bacteriol.* 192, 6428–6438.
- Dobson, P.D., Smallbone, K., Jameson, D., Simeonidis, E., Lanthaler, K., Pir, P., Lu, C., Swainston, N., Dunn, W.B., Fisher, P., et al. (2010). Further developments towards a genome-scale metabolic model of yeast. *BMC Syst. Biol.* 4, 145.
- Esnault, C., Dulermo, T., Smirnov, A., Askara, A., David, M., Deniset-Besseau, A., Holland, I.-B., and Virolle, M.-J. (2017). Strong antibiotic production is correlated with highly active oxidative metabolism in *Streptomyces coelicolor* M145. *Sci. Rep.* 7, 200.
- Feist, A.M., Henry, C.S., Reed, J.L., Krummenacker, M., Joyce, A.R., Karp, P.D., Broadbelt, L.J., Hatzimanikatis, V., and Palsson, B.O. (2007). A genome-scale metabolic reconstruction for *Escherichia coli* K-12 MG1655 that accounts for 1260 ORFs and thermodynamic information. *Mol. Syst. Biol.* 3, 121.
- Fink, D., Weißschuh, N., Reuther, J., Wohlleben, W., and Engels, A. (2002). Two transcriptional regulators GlnR and GlnRII are involved in regulation of nitrogen metabolism in *Streptomyces coelicolor* A3(2). *Mol. Microbiol.* 46, 331–347.
- Flamholz, A., Noor, E., Bar-Even, A., and Milo, R. (2012). eQuilibrator—the biochemical thermodynamics calculator. *Nucleic Acids Res.* 40, D770–D775.
- Gaal, T., Bartlett, M.S., Ross, W., Turnbough, C.L., and Gourse, R.L. (1997). Transcription regulation by initiating NTP concentration: rRNA synthesis in bacteria. *Science* 278, 2092–2097.
- Gallo, G., Renzone, G., Alduina, R., Stegmann, E., Weber, T., Lantz, A.E., Thykaer, J., Sangiorgi, F., Scaloni, A., and Puglia, A.M. (2010). Differential proteomic analysis reveals novel links between primary metabolism and antibiotic production in *Amycolatopsis balhimycina*. *Proteomics* 10, 1336–1358.
- Getsin, I., Nalbandian, G.H., Yee, D.C., Vastermark, A., Paparoditis, P.C., Reddy, V.S., and Saier, M.H. (2013). Comparative genomics of transport proteins in developmental bacteria: *myxococcus xanthus* and *Streptomyces coelicolor*. *BMC Microbiol.* 13, 279.
- Gomez-Escribano, J.P., and Bibb, M.J. (2011). Engineering *Streptomyces coelicolor* for heterologous expression of secondary metabolite gene clusters. *Microb. Biotechnol.* 4, 207–215.
- Gomez-Escribano, J.P., and Bibb, M.J. (2014). Heterologous expression of natural product biosynthetic gene clusters in *Streptomyces coelicolor*: from genome mining to manipulation of biosynthetic pathways. *J. Ind. Microbiol. Biotechnol.* 41, 425–431.
- Gomez-Escribano, J.P., Song, L., Fox, D.J., Yeo, V., Bibb, M.J., and Challis, G.L. (2012). Structure and biosynthesis of the unusual polyketide alkaloid coelimycin P1, a metabolic product of the cpk gene cluster of *Streptomyces coelicolor* M145. *Chem. Sci.* 3, 2716–2720.
- Gu, C., Kim, G.B., Kim, W.J., Kim, H.U., and Lee, S.Y. (2019). Current status and applications of genome-scale metabolic models. *Genome Biol.* 20, 121.
- Hahn, J.-S., Oh, S.-Y., and Roe, J.-H. (2002). Role of OxyR as a peroxide-sensing positive regulator in *Streptomyces coelicolor* A3(2). *J. Bacteriol.* 184, 5214–5222.
- Hastings, J., Owen, G., Dekker, A., Ennis, M., Kale, N., Muthukrishnan, V., Turner, S., Swainston, N., Mendes, P., and Steinbeck, C. (2016). ChEBI in 2016: improved services and an expanding collection of metabolites. *Nucleic Acids Res.* 44, D1214–D1219.
- Heavner, B.D., Smallbone, K., Barker, B., Mendes, P., and Walker, L.P. (2012). Yeast 5—an expanded reconstruction of the *Saccharomyces cerevisiae* metabolic network. *BMC Syst. Biol.* 6, 55.
- Heavner, B.D., Smallbone, K., Price, N.D., and Walker, L.P. (2013). Version 6 of the consensus yeast metabolic network refines biochemical coverage and improves model performance. *Database* 2013, bat059.
- Hefzi, H., Ang, K.S., Hanscho, M., Bordbar, A., Ruckerbauer, D., Lakshmanan, M., Orellana, C.A., Baycin-Hizal, D., Huang, Y., Ley, D., et al. (2016). A consensus genome-scale reconstruction of Chinese hamster ovary cell metabolism. *Cell Syst.* 3, 434–443.e8.
- Herrgard, M.J., Swainston, N., Dobson, P., Dunn, W.B., Arga, K.Y., Arvas, M., Blüthgen, N., Borger, S., Costenoble, R., Heinemann, M., et al. (2008). A consensus yeast metabolic network reconstruction obtained from a community approach to systems biology. *Nat. Biotechnol.* 26, 1155–1160.
- Hesketh, A., Chen, W.J., Ryding, J., Chang, S., and Bibb, M. (2007). The global role of ppGpp synthesis in morphological differentiation and antibiotic production in *Streptomyces coelicolor* A3(2). *Genome Biol.* 8, R161.
- Horinouchi, S. (2003). AfsR as an integrator of signals that are sensed by multiple serine/threonine kinases in *Streptomyces coelicolor* A3(2). *J. Ind. Microbiol. Biotechnol.* 30, 462–467.
- Hu, H., Zhang, Q., and Ochi, K. (2002). Activation of antibiotic biosynthesis by specified mutations in the rpoB gene (encoding the RNA polymerase β subunit) of *Streptomyces lividans*. *J. Bacteriol.* 184, 3984–3991.
- Hutchings, M.I., Hoskisson, P.A., Chandra, G., and Buttner, M.J. (2004). Sensing and responding to diverse extracellular signals? Analysis of the sensor kinases and response regulators of *Streptomyces coelicolor* A3(2). *Microbiology* 150, 2795–2806.
- Jager, G., Battke, F., and Nieselt, K. (2011). Tiala—time series alignment analysis. In 2011 IEEE Symposium on Biological Data Visualization (BioVis), pp. 55–61.
- Jones, G., Sol, R.D., Dudley, E., and Dyson, P. (2011). Forkhead-associated proteins genetically linked to the serine/threonine kinase PknB regulate carbon flux towards antibiotic

- biosynthesis in *Streptomyces coelicolor*. *Microb. Biotechnol.* 4, 263–274.
- Jonsbu, E., Christensen, B., and Nielsen, J. (2001). Changes of in vivo fluxes through central metabolic pathways during the production of nystatin by *Streptomyces noursei* in batch culture. *Appl. Microbiol. Biotechnol.* 56, 93–100.
- Kanehisa, M. (2000). KEGG: kyoto encyclopedia of genes and genomes. *Nucleic Acids Res.* 28, 27–30.
- Kanehisa, M., Sato, Y., Furumichi, M., Morishima, K., and Tanabe, M. (2019). New approach for understanding genome variations in KEGG. *Nucleic Acids Res.* 47, D590–D595.
- Karp, P.D., Billington, R., Caspi, R., Fulcher, C.A., Latendresse, M., Kothari, A., Keseler, I.M., Krummenacker, M., Midford, P.E., Ong, Q., et al. (2019). The BioCyc collection of microbial genomes and metabolic pathways. *Brief. Bioinform.* 20, 1085–1093.
- Kepplinger, B., Morton-Laing, S., Seistrup, K.H., Marrs, E.C.L., Hopkins, A.P., Perry, J.D., Strahl, H., Hall, M.J., Errington, J., and Allenby, N.E.E. (2018). Mode of action and heterologous expression of the natural product antibiotic vancomycin. *ACS Chem. Biol.* 13, 207–214.
- Kieser, T., Bibb, M.J., Buttner, M.J., Chater, K.F., and Hopwood, D.A. (2000). *Practical Streptomyces Genetics* (Norwich, UK: John Innes Foundation).
- Kim, D.-J., Huh, J.-H., Yang, Y.-Y., Kang, C.-M., Lee, I.-H., Hyun, C.-G., Hong, S.-K., and Suh, J.-W. (2018). Accumulation of S-Adenosyl-L-Methionine enhances production of actinorhodin but inhibits sporulation in *streptomyces lividans* TK23. *J. Bacteriol.* 185, 592–600.
- Kim, M.W., Sang Yi, J., Kim, J.-N.J.J.-N.N.J., Kim, J.-N.J.J.-N.N.J., Kim, M.W., and Kim, B.-G.G. (2014). Reconstruction of a high-quality metabolic model enables the identification of gene overexpression targets for enhanced antibiotic production in *streptomyces coelicolor* A3(2). *Biotechnol. J.* 9, 1185–1194.
- King, Z.A., Dräger, A., Ebrahim, A., Sonnenschein, N., Lewis, N.E., and Palsson, B.O. (2015). Escher: a web application for building, sharing, and embedding data-Rich visualizations of biological pathways. *PLOS Comput. Biol.* 11, e1004321.
- King, Z.A., Lu, J., Dräger, A., Miller, P., Federowicz, S., Lerman, J.A., Ebrahim, A., Palsson, B.O., and Lewis, N.E. (2016). BiGG Models: a platform for integrating, standardizing and sharing genome-scale models. *Nucleic Acids Res.* 44, D515–D522.
- Kumelj, T., Sulheim, S., Wentzel, A., and Almaas, E. (2019). Predicting strain engineering strategies using iKS1317: a genome-scale metabolic model of *Streptomyces coelicolor*. *Biotechnol. J.* 14, 1800180.
- Lamb, D.C., Ikeda, H., Nelson, D.R., Ishikawa, J., Skaug, T., Jackson, C., Omura, S., Waterman, M.R., and Kelly, S.L. (2003). Cytochrome P450 complement (CYPome) of the avermectin-producer *Streptomyces avermitilis* and comparison to that of *Streptomyces coelicolor* A3(2). *Biochem. Biophys. Res. Commun.* 307, 610–619.
- Latifi, A., Ruiz, M., and Zhang, C.-C. (2009). Oxidative stress in cyanobacteria. *FEMS Microbiol. Rev.* 33, 258–278.
- Lee, P.-C., Umeyama, T., and Horinouchi, S. (2002). afsS is a target of AfsR, a transcriptional factor with ATPase activity that globally controls secondary metabolism in *Streptomyces coelicolor* A3(2). *Mol. Microbiol.* 43, 1413–1430.
- Li, T., Du, Y., Cui, Q., Zhang, J., Zhu, W., Hong, K., and Li, W. (2013). Cloning, characterization and heterologous expression of the indolocarbazole biosynthetic gene cluster from marine-derived streptomycetes *saneyensis* FMA. *Mar. Drugs* 11, 466–488.
- Li, X., Wang, J., Li, S., Ji, J., Wang, W., and Yang, K. (2015). ScbR- and ScbR2-mediated signal transduction networks coordinate complex physiological responses in *Streptomyces coelicolor*. *Sci. Rep.* 5, 14831.
- Li, L., Jiang, W., and Lu, Y. (2017). A novel two-component system, GluR-GluK, involved in glutamate sensing and uptake in *streptomyces coelicolor*. *J. Bacteriol.* 199, e00097-17.
- Liao, Y., Smyth, G.K., and Shi, W. (2014). featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. *Bioinformatics* 30, 923–930.
- Licona-Cassani, C., Marcellin, E., Quek, L.-E., Jacob, S., and Nielsen, L.K. (2012). Reconstruction of the *Saccharopolyspora erythraea* genome-scale model and its use for enhancing erythromycin production. *Antonie van Leeuwenhoek* 102, 493–502.
- Lieven, C., Beber, M.E., Olivier, B.G., Bergmann, F.T., Ataman, M., Babaei, P., Bartell, J.A., Blank, L.M., Chauhan, S., Correia, K., et al. (2020). MEMOTE for standardized genome-scale metabolic model testing. *Nat. Biotechnol.* 38, 272–276.
- Liu, G., Chater, K.F., Chandra, G., Niu, G., and Tan, H. (2013). Molecular regulation of antibiotic biosynthesis in *streptomyces*. *Microbiol. Mol. Biol. Rev.* 77, 112–143.
- Love, M.I., Huber, W., and Anders, S. (2014). Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol.* 15, 550.
- Lu, H., Li, F., Sánchez, B.J., Zhu, Z., Li, G., Domenzain, I., Marcisauskas, S., Anton, P.M., Lappa, D., Lieven, C., et al. (2019). A consensus *S. cerevisiae* metabolic model Yeast8 and its ecosystem for comprehensively probing cellular metabolism. *Nat. Commun.* 10, 1–13.
- Martin, J.F. (2004). Phosphate control of the biosynthesis of antibiotics and other secondary metabolites is mediated by the PhoR-PhoP system: an unfinished story. *J. Bacteriol.* 186, 5197–5201.
- Martin, J.F., Santos-Beneit, F., Rodríguez-García, A., Sola-Landa, A., Smith, M.C.M., Ellingsen, T.E., Nieselt, K., Burroughs, N.J., and Wellington, E.M.H. (2012). Transcriptomic studies of phosphate control of primary and secondary metabolism in *Streptomyces coelicolor*. *Appl. Microbiol. Biotechnol.* 95, 61–75.
- Martin, J.F., Rodríguez-García, A., and Liras, P. (2017). The master regulator PhoP coordinates phosphate and nitrogen metabolism, respiration, cell differentiation and antibiotic biosynthesis: comparison in *Streptomyces coelicolor* and *Streptomyces avermitilis*. *J. Antibiot.* 70, 534–541.
- Martin-Martin, S., Rodríguez-García, A., Santos-Beneit, F., Franco-Dominguez, E., Sola-Landa, A., and Martin, J.F. (2018). Self-control of the PHO regulon: the PhoP-dependent protein PhoU controls negatively expression of genes of PHO regulon in *Streptomyces coelicolor*. *J. Antibiot.* 71, 113–122.
- Mi, H., Muruganujan, A., Ebert, D., Huang, X., and Thomas, P.D. (2019). PANTHER version 14: more genomes, a new PANTHER GO-slim and improvements in enrichment analysis tools. *Nucleic Acids Res.* 47, D419–D426.
- Millan-Oropeza, A., Henry, C., Blein-Nicolas, M., Aubert-Frambourg, A., Moussa, F., Bleton, J., and Virolle, M.-J. (2017). Quantitative proteomics analysis confirmed oxidative metabolism predominates in *streptomyces coelicolor* versus glycolytic metabolism in *streptomyces lividans*. *J. Proteome Res.* 16, 2597–2613.
- Mohite, O.S., Weber, T., Kim, H.U., and Lee, S.Y. (2019). Genome-scale metabolic reconstruction of actinomycetes for antibiotics production. *Biotechnol. J.* 14, 1800377.
- Moretti, S., Martin, O., Van Du Tran, T., Bridge, A., Morgat, A., and Pagni, M. (2016). MetaNetX/MNXref – reconciliation of metabolites and biochemical reactions to bring together genome-scale metabolic networks. *Nucleic Acids Res.* 44, D523–D526.
- Naseer, N., Shapiro, J.A., and Chander, M. (2014). RNA-Seq analysis reveals a six-gene SoxR regulon in *Streptomyces coelicolor*. *PLoS One* 9, e106181.
- Nepal, K.K., and Wang, G. (2019). *Streptomyces*: surrogate hosts for the genetic manipulation of biosynthetic gene clusters and production of natural products. *Biotechnol. Adv.* 37, 1–20.
- Nieselt, K., Battke, F., Herbig, A., Bruheim, P., Wentzel, A., Jakobsen, Ø.M., Sletta, H., Alam, M.T., Merlo, M.E., Moore, J., et al. (2010). The dynamic architecture of the metabolic switch in *Streptomyces coelicolor*. *BMC Genomics* 11, 10.
- Nothhaft, H., Rigali, S., Boomsma, B., Swiatek, M., McDowall, K.J., van Wezel, G.P., and Titgemeyer, F. (2010). The permease gene nagE2 is the key to N-acetylglucosamine sensing and utilization in *Streptomyces coelicolor* and is subject to multi-level control. *Mol. Microbiol.* 75, 1133–1144.
- Okamoto, S., Lezhava, A., Hosaka, T., Okamoto-Hosoya, Y., and Ochi, K. (2003). Enhanced expression of S-adenosylmethionine synthetase causes overproduction of actinorhodin in *streptomyces coelicolor* A3(2). *J. Bacteriol.* 185, 601–609.
- Orth, J.D., Thiele, I., and Palsson, B.Ø.O. (2010). What is flux balance analysis? *Nat. Biotech.* 28, 245–248.
- Perez-Riverol, Y., Csordas, A., Bai, J., Bernal-Llinares, M., Hewapathirana, S., Kundu, D.J., Inuganti, A., Griss, J., Mayer, G., Eisenacher, M.,

- et al. (2019). The PRIDE database and related tools and resources in 2019: improving support for quantification data. *Nucleic Acids Res.* **47**, D442–D450.
- Rigali, S., Titgemeyer, F., Barends, S., Mulder, S., Thomae, A.W., Hopwood, D.A., and van Wezel, G.P. (2008). Feast or famine: the global regulator DasR links nutrient stress to antibiotic production by *Streptomyces*. *EMBO Rep.* **9**, 670–675.
- Robinson, J.L., and Nielsen, J. (2016). Integrative analysis of human omics data using biomolecular networks. *Mol. Biosyst.* **12**, 2953–2964.
- Rodríguez, E., Navone, L., Casati, P., and Gramajo, H. (2012). Impact of malic enzymes on antibiotic and triacylglycerol production in *Streptomyces coelicolor*. *Appl. Environ. Microbiol.* **78**, 4571–4579.
- Rutledge, P.J., and Challis, G.L. (2015). Discovery of microbial natural products by activation of silent biosynthetic gene clusters. *Nat. Rev. Microbiol.* **13**, 509–523.
- Sanchez, B.J., Zhang, C., Nilsson, A., Lahtvee, P.-J., Kerkhoven, E.J., and Nielsen, J. (2017). Improving the phenotype predictions of a yeast genome-scale metabolic model by incorporating enzymatic constraints. *Mol. Syst. Biol.* **13**, 935.
- Shin, J.-H., Singh, A.K., Cheon, D.-J., and Roe, J.-H. (2011). Activation of the SoxR regulon in *Streptomyces coelicolor* by the extracellular form of the pigmented antibiotic actinorhodin. *J. Bacteriol.* **193**, 75–81.
- Smirnov, A., Esnault, C., Prigent, M., Holland, I.B., and Virolle, M.-J. (2015). Phosphate homeostasis in conditions of phosphate proficiency and limitation in the wild type and the *phoP* mutant of *Streptomyces lividans*. *PLoS One* **10**, e0126221.
- Sola-Landa, A., Moura, R.S., and Martin, J.F. (2003). The two-component PhoR-PhoP system controls both primary metabolism and secondary metabolite biosynthesis in *Streptomyces lividans*. *Proc. Natl. Acad. Sci. U S A* **100**, 6133–6138.
- Sola-Landa, A., Rodríguez-García, A., Franco-Domínguez, E., and Martin, J.F. (2005). Binding of PhoP to promoters of phosphate-regulated genes in *Streptomyces coelicolor*: identification of PHO boxes. *Mol. Microbiol.* **56**, 1373–1385.
- Srivatsan, A., and Wang, J.D. (2008). Control of bacterial transcription, translation and replication by (p)ppGpp. *Curr. Opin. Microbiol.* **11**, 100–105.
- Stahl, W., and Sies, H. (2003). Antioxidant activity of carotenoids. *Mol. Aspects Med.* **24**, 345–351.
- Stirrett, K., Denoya, C., and Westpheling, J. (2009). Branched-chain amino acid catabolism provides precursors for the Type II polyketide antibiotic, actinorhodin, via pathways that are nutrient dependent. *J. Ind. Microbiol. Biotechnol.* **36**, 129–137.
- Thanapitsiri, A., Claesen, J., Gomez-Escribano, J.-P., Bibb, M., and Thamchaipenet, A. (2015). A *Streptomyces coelicolor* host for the heterologous expression of Type III polyketide synthase genes. *Microb. Cell Fact.* **14**, 145.
- The Gene Ontology Consortium (2019). The gene ontology resource: 20 years and still GOing strong. *Nucleic Acids Res.* **47**, D330–D338.
- The UniProt Consortium (2019). UniProt: a worldwide hub of protein knowledge. *Nucleic Acids Res.* **47**, D506–D515.
- Thiele, I., and Palsson, B.Ø. (2010). A protocol for generating a high-quality genome-scale metabolic reconstruction. *Nat. Protoc.* **5**, 93–121.
- Thiele, I., Swainston, N., Fleming, R.M.T., Hoppe, A., Sahoo, S., Aurich, M.K., Haraldsdottir, H., Mo, M.L., Rolfsson, O., Stobbe, M.D., et al. (2013). A community-driven global reconstruction of human metabolism. *Nat. Biotechnol.* **31**, 419–425.
- Thomas, L., Hodgson, D.A., Wentzel, A., Nieselt, K., Ellingsen, T.E., Moore, J., Morrissey, E.R., Legaie, R., STREAM Consortium, T.S., Wohlleben, W., et al. (2012). Metabolic switches and adaptations deduced from the proteomes of *Streptomyces coelicolor* wild type and *phoP* mutant grown in batch culture. *Mol. Cell. Proteomics* **11**, M111.013797.
- Toro, L., Pinilla, L., Avignone-Rossa, C., and Ríos-Estapa, R. (2018). An enhanced genome-scale metabolic reconstruction of *Streptomyces clavuligerus* identifies novel strain improvement strategies. *Bioproc. Biosyst. Eng.* **41**, 657–669.
- Valton, J., Mathevon, C., Fontecave, M., Nivière, V., and Ballou, D.P. (2008). Mechanism and regulation of the two-component FMN-dependent monoxygenase ActVA-ActVB from *Streptomyces coelicolor*. *J. Biol. Chem.* **283**, 10287–10296.
- Valverde, J.R., Gullón, S., and Mellado, R.P. (2018). Modelling the metabolism of protein secretion through the Tat route in *Streptomyces lividans*. *BMC Microbiol.* **18**, 59.
- Wang, H., Marcišauskas, S., Sánchez, B.J., Domenzain, I., Hermansson, D., Agren, R., Nielsen, J., and Kerkhoven, E.J. (2018). Raven 2.0: a versatile toolbox for metabolic network reconstruction and a case study on *Streptomyces coelicolor*. *PLOS Comput. Biol.* **14**, e1006541.
- Wentzel, A., Bruheim, P., Øverby, A., Jakobsen, Ø.M., Sletta, H., Omara, W.A.M., Hodgson, D.A., and Ellingsen, T.E. (2012a). Optimized submerged batch fermentation strategy for systems scale studies of metabolic switching in *Streptomyces coelicolor* A3(2). *BMC Syst. Biol.* **6**, 59.
- Wentzel, A., Sletta, H., Consortium, S., Ellingsen, T.E., and Bruheim, P. (2012b). Intracellular metabolite pool changes in response to nutrient depletion induced metabolic switching in *Streptomyces coelicolor*. *Metabolites* **2**, 178–194.
- Xu, J., Tozawa, Y., Lai, C., Hayashi, H., and Ochi, K. (2002). A rifampicin resistance mutation in the *rpoB* gene confers ppGpp-independent antibiotic production in *Streptomyces coelicolor* A3(2). *Mol. Gen. Genomics* **268**, 179–189.
- Yin, J., Hoffmann, M., Bian, X., Tu, Q., Yan, F., Xia, L., Ding, X., Francis Stewart, A., Müller, R., Fu, J., et al. (2015). Direct cloning and heterologous expression of the salinomycin biosynthetic gene cluster from *Streptomyces albus* DSM41398 in *Streptomyces coelicolor* A3(2). *Sci. Rep.* **5**, 15081.
- Yoshida, M., Muneyuki, E., and Hisabori, T. (2001). ATP synthase — a marvellous rotary engine of the cell. *Nat. Rev. Mol. Cell Biol.* **2**, 669–677.
- Zangar, R.C., Davydov, D.R., and Verma, S. (2004). Mechanisms that regulate production of reactive oxygen species by cytochrome P450. *Toxicol. Appl. Pharmacol.* **199**, 316–331.

iScience, Volume 23

Supplemental Information

Enzyme-Constrained Models and Omics Analysis of *Streptomyces coelicolor* Reveal Metabolic Changes that Enhance Heterologous Production

Snorre Sulheim, Tjaša Kumelj, Dino van Dissel, Ali Salehzadeh-Yazdi, Chao Du, Gilles P. van Wezel, Kay Nieselt, Eivind Almaas, Alexander Wentzel, and Eduard J. Kerkhoven

Supplemental figures

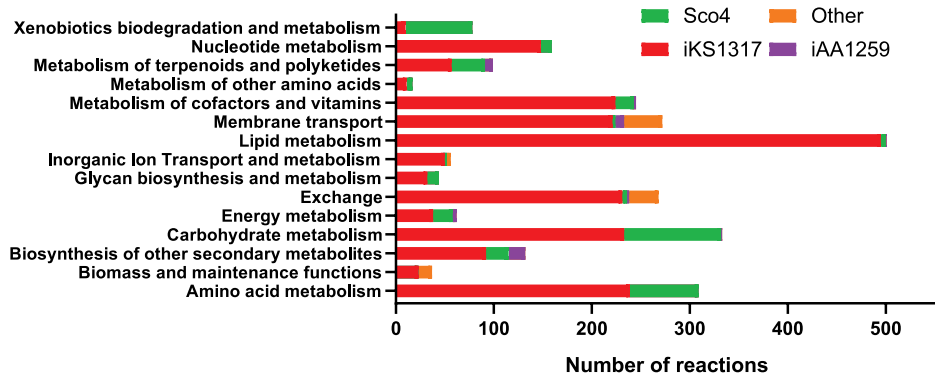


Figure S1: Reaction subsystems and origin, related to Figure 1A. The number of reactions in Sco-GEM in each of the 15 subsystems, and from which model they originate from. The other reactions (orange) are added during reconstruction of Sco-GEM.

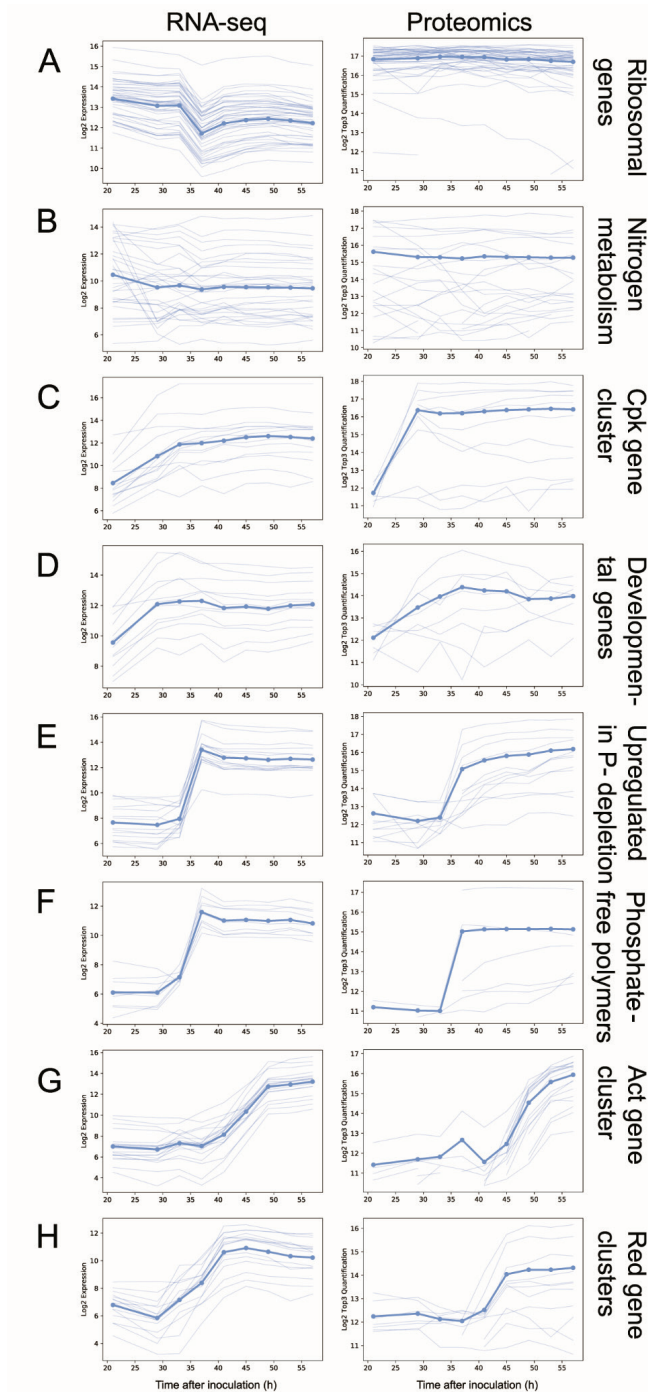


Figure S2: Gene clusters associated with metabolic switch, related to Figure 2C. RNA-seq (left column) and proteomics (right column) from M145 of the 8 gene clusters associated with the metabolic switch as previously identified (Nieselt et al., 2010). The 8 clusters are: A) genes related to ribosomal proteins; B) genes related to nitrogen metabolism; C) Cpk gene cluster; D) genes related to development; E) genes upregulated in response to phosphate depletion; F) genes involved in synthesis of phosphate-free polymers; G) Act gene cluster; H) Red gene cluster.

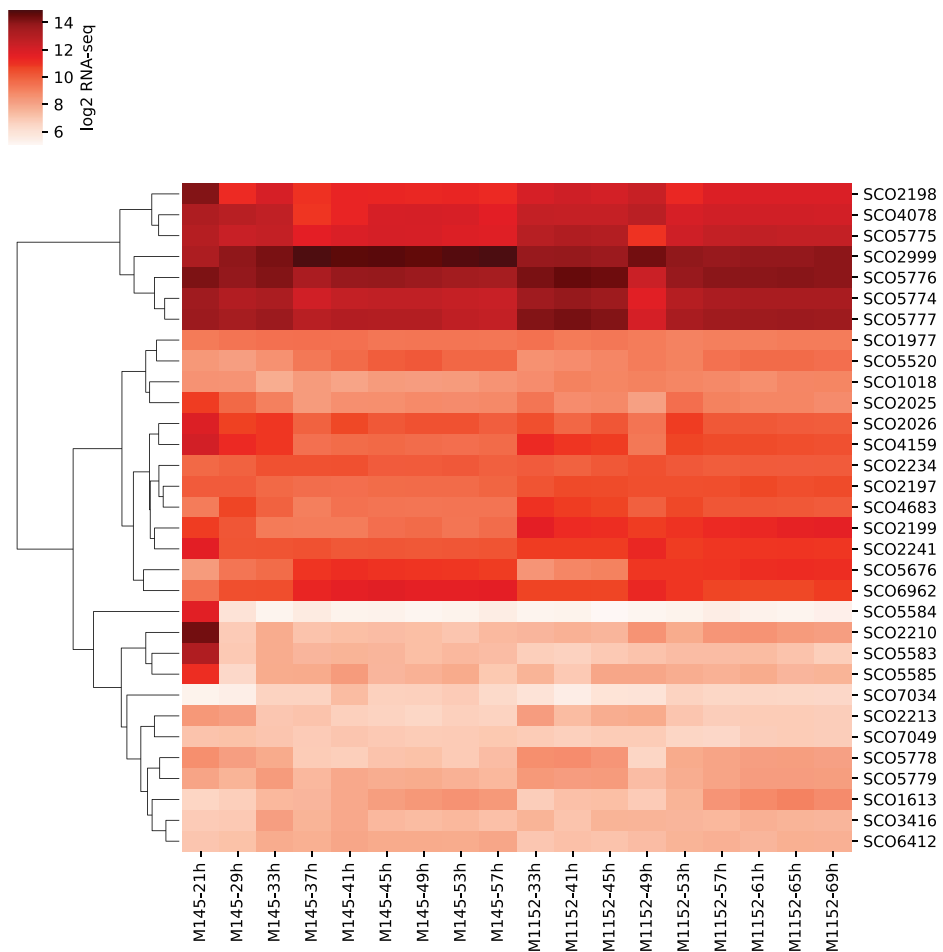


Figure S3: Log-transformed expression levels of genes associated with nitrogen metabolism, related to Figure 2D. The order of the genes is determined by hierarchical clustering to align genes with similar expression profiles next to each other. From the log₂-transformed RNA-seq data we observe that glutamate import (SCO5774-5777), the glutamate sensing system *gluR-gluK* (SCO5778 and SCO5779), *glnR* (SCO4159) and *glnA* (SCO2198) are downregulated subsequent to phosphate depletion. The phosphate depletion occurs between the third and fourth time point, i.e. at 35 and 47 hours for M145 and M1152, respectively. We also observe that the first time point in M145 is very different from all other samples.

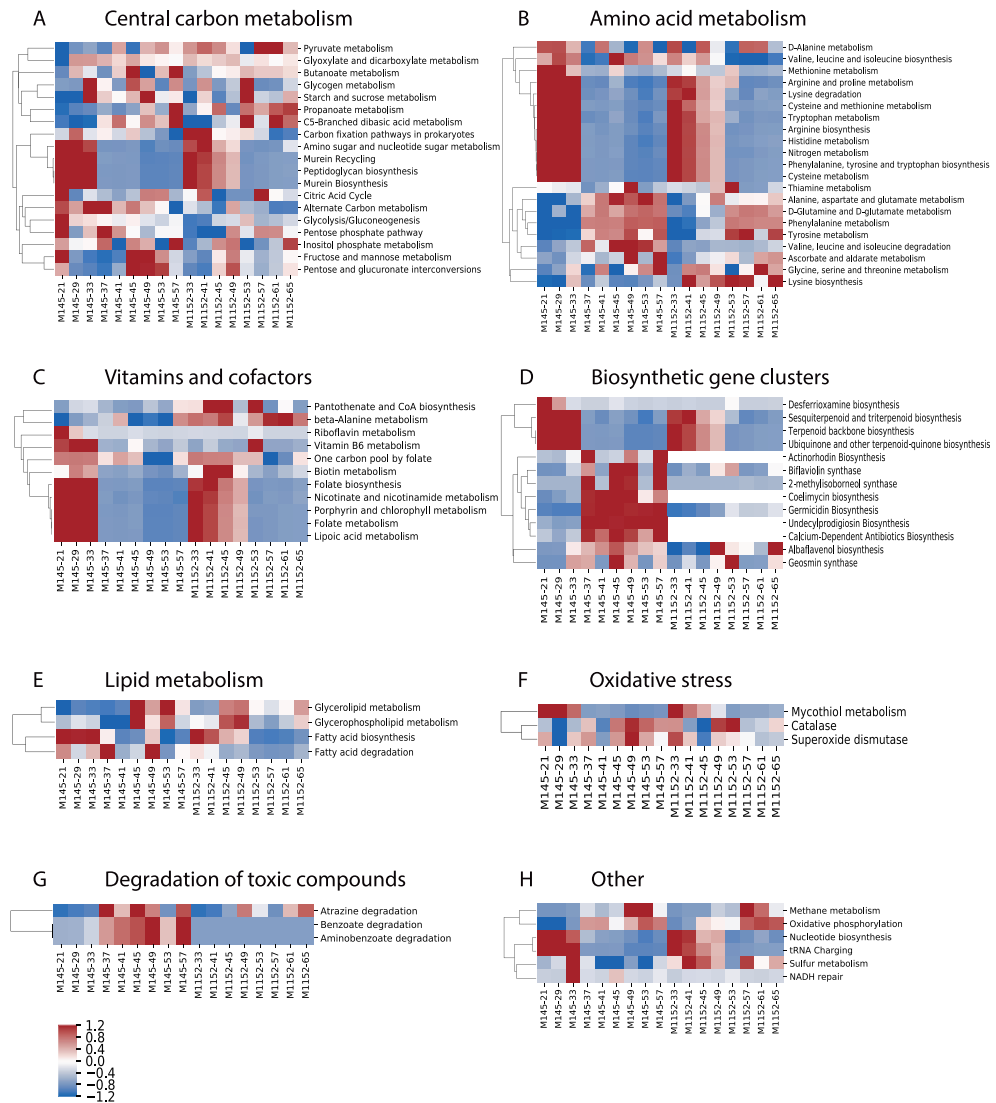


Figure S4: Clustered heatmaps of Z-score based on CO₂-normalized sum of fluxes of all pathways standardized within each pathway and separated into different subsystems / parts of the metabolism. Related to Figure 2D. A) Central carbon metabolism. B) Amino acid metabolism. C) Metabolism of vitamins and cofactors. D) Pathways of Biosynthetic gene clusters. E) Lipid metabolism. F) Oxidative stress. G) Degradation of toxic compounds. H) All other pathways. For all panels only pathway with a minimum flux of 1e-8 mmol (g DW)⁻¹ h⁻¹ were included.

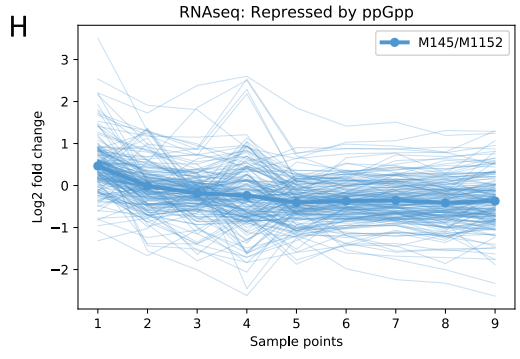
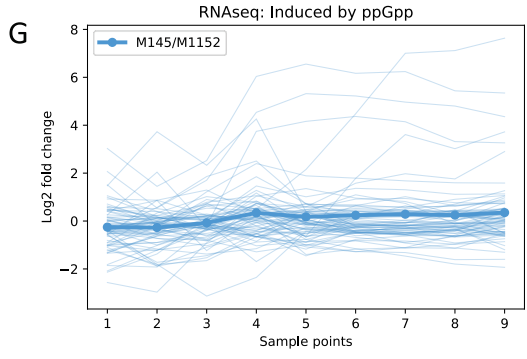
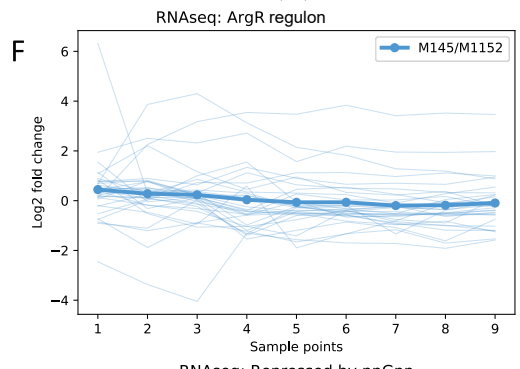
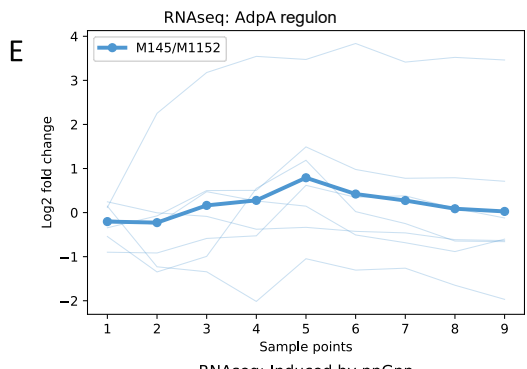
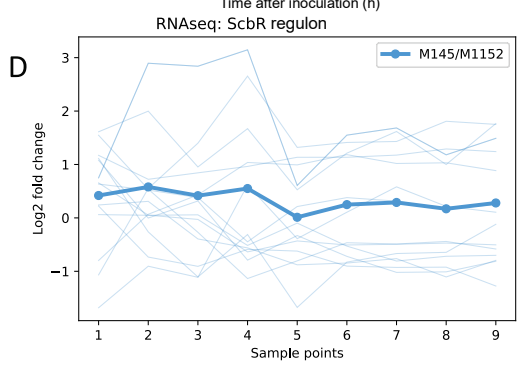
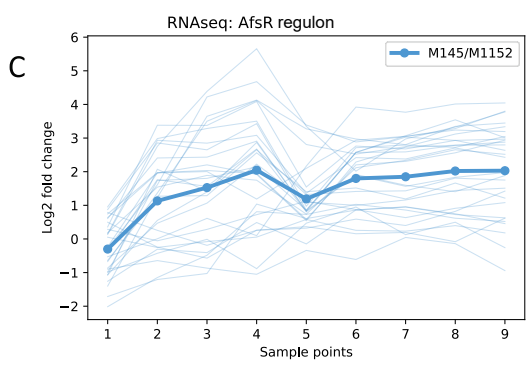
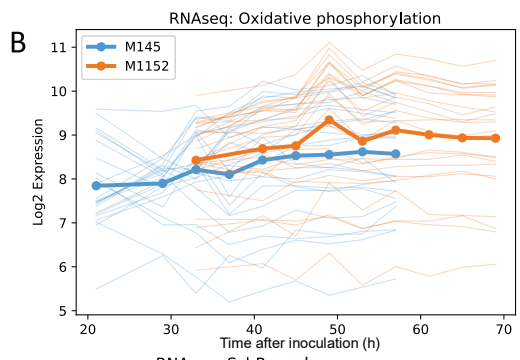
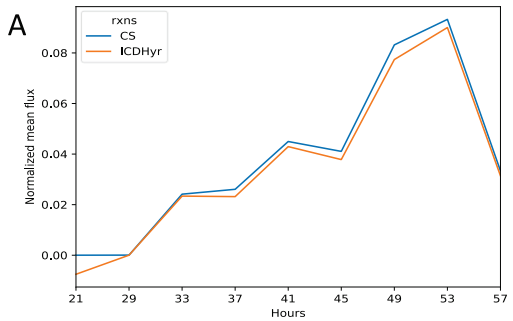


Figure S5: RNA-seq, proteome and flux prediction of specific gene clusters and reactions.

Related to Figure 2, 3 and S7. A) This panel display increasing CO₂-normalized flux through citrate synthase (CS) and isocitrate dehydrogenase (ICDHyr) at later time points in M145 as predicted by EcSco-GEM. These two reactions are both part of the TCA cycle, converting acetyl-CoA to citrate (citrate synthase) and isocitrate to alpha-ketogluterate (isocitrate dehydrogenase). B) Log₂ normalized expression data of genes involved in oxidative phosphorylation for M145 (blue) and M1152 (orange). The average expression level is higher in M1152 than in M145 but increasing at later time points for both strains. The expression profiles are only partially overlapping along the x-axis (hours after inoculation) because of the reduced growth and therefore delayed cultivation of M1152. C-H) Comparison of log₂ normalized expression data as calculated with $(\log_2 M145) - \log_2(M1152)$, where positive values indicate upregulation in M145 relative to M1152, and vice versa for negative values. C) Increased expression of genes of the AfsR regulon in M145, while no significant difference in expression is observed for (D) ScbR regulon; (E) AdpA regulon; (F) ArgR regulon; (G) genes induced by ppGpp; and (H) genes repressed by ppGpp.

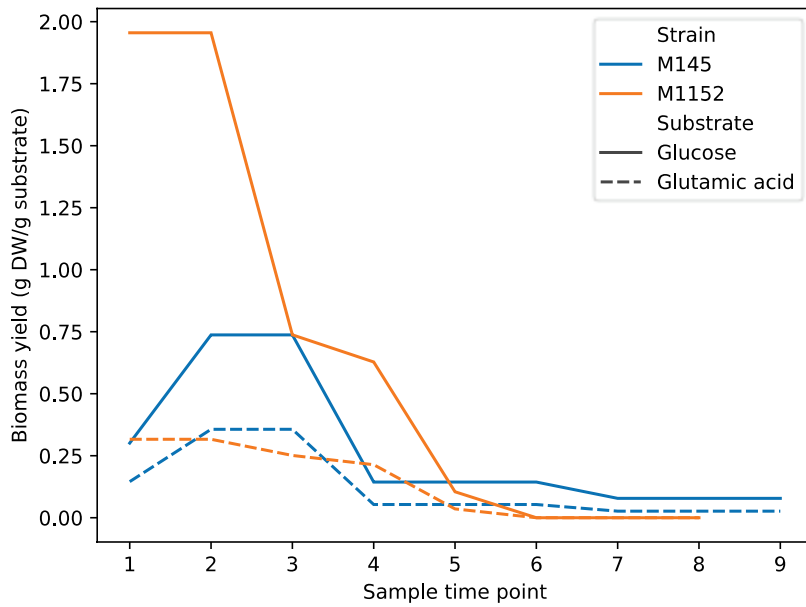


Figure S6: Biomass yield on glucose and glutamic acid, related to Figure 4. M1152 (orange) has a higher growth yield on glucose than M145 (blue). The yield on glutamic acid (dashed line) is similar between the two strains.

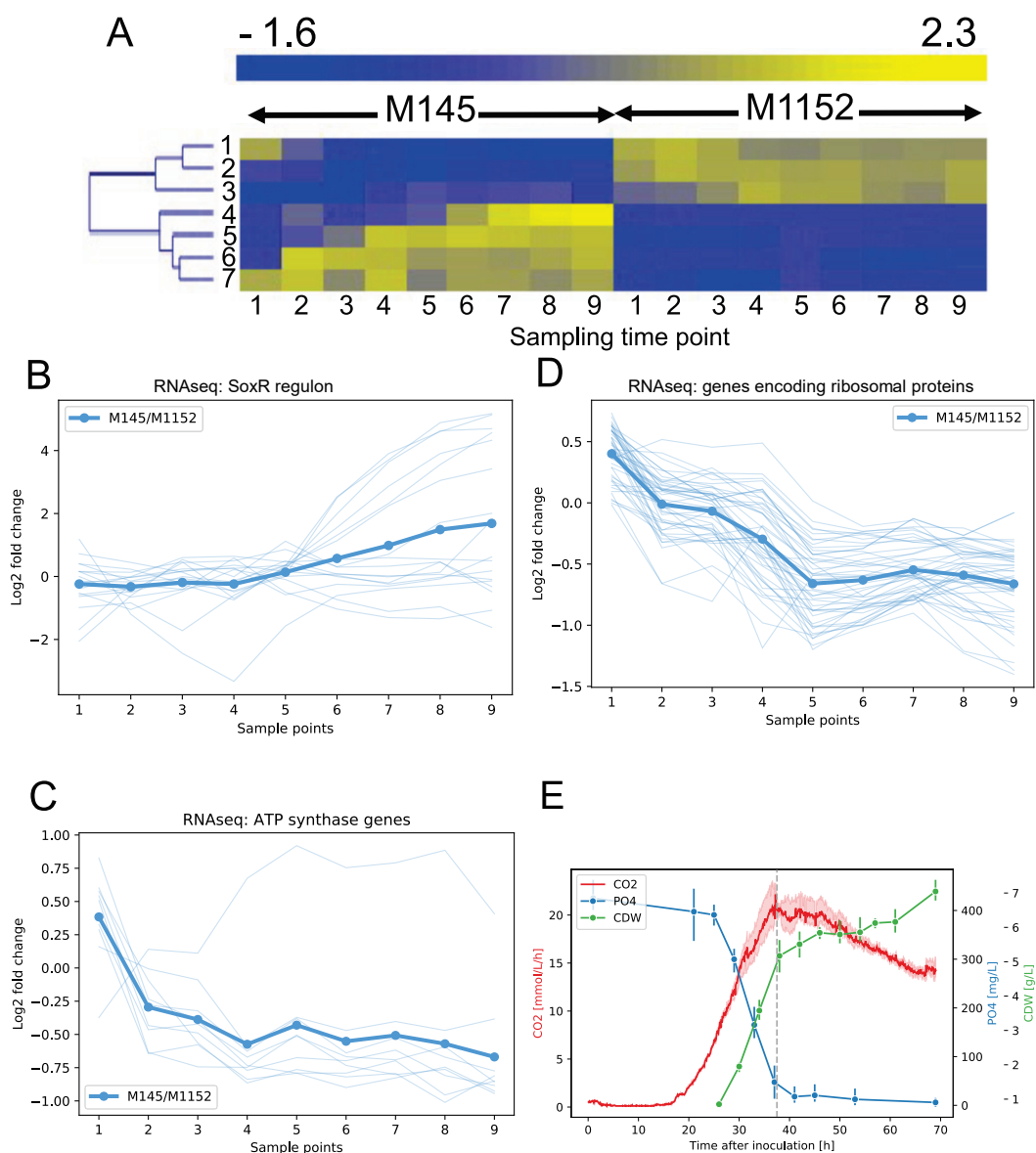


Figure S7: Analysis of transcriptome data of genes, related to Figure 2, 3, S5 and S8, and cultivation data of M1146, related to Figure 7. A) The heatmap display the mean standardized log₂ expression levels for the 7 clusters of differentially expressed genes as determined by unsupervised clustering (k-means). Cluster 1-3 are upregulated in M1152, while the last four

(cluster 4-7) are upregulated from the beginning or at later time points in M145. B-D) Comparison of log₂ normalized expression data as calculated with $(\log_2 M145) - \log_2(M1152)$, where positive values indicate upregulation in M145 relative to M1152, and vice versa for negative values. B) Genes in the SoxR regulon are reducing expression in M1152 at later time points. C) Almost all genes in the ATP-synthase cluster are up-regulated in M1152 after the first time point. D) The transcription of ribosomal protein genes after the metabolic switch is increased in M1152 compared to M145. E) Batch cultivation data of *S. coelicolor* M1146, showing volume corrected respiration (CO₂), phosphate (PO₄) and cell dry weight (CDW). Error bars are standard deviations of three biological replicates.

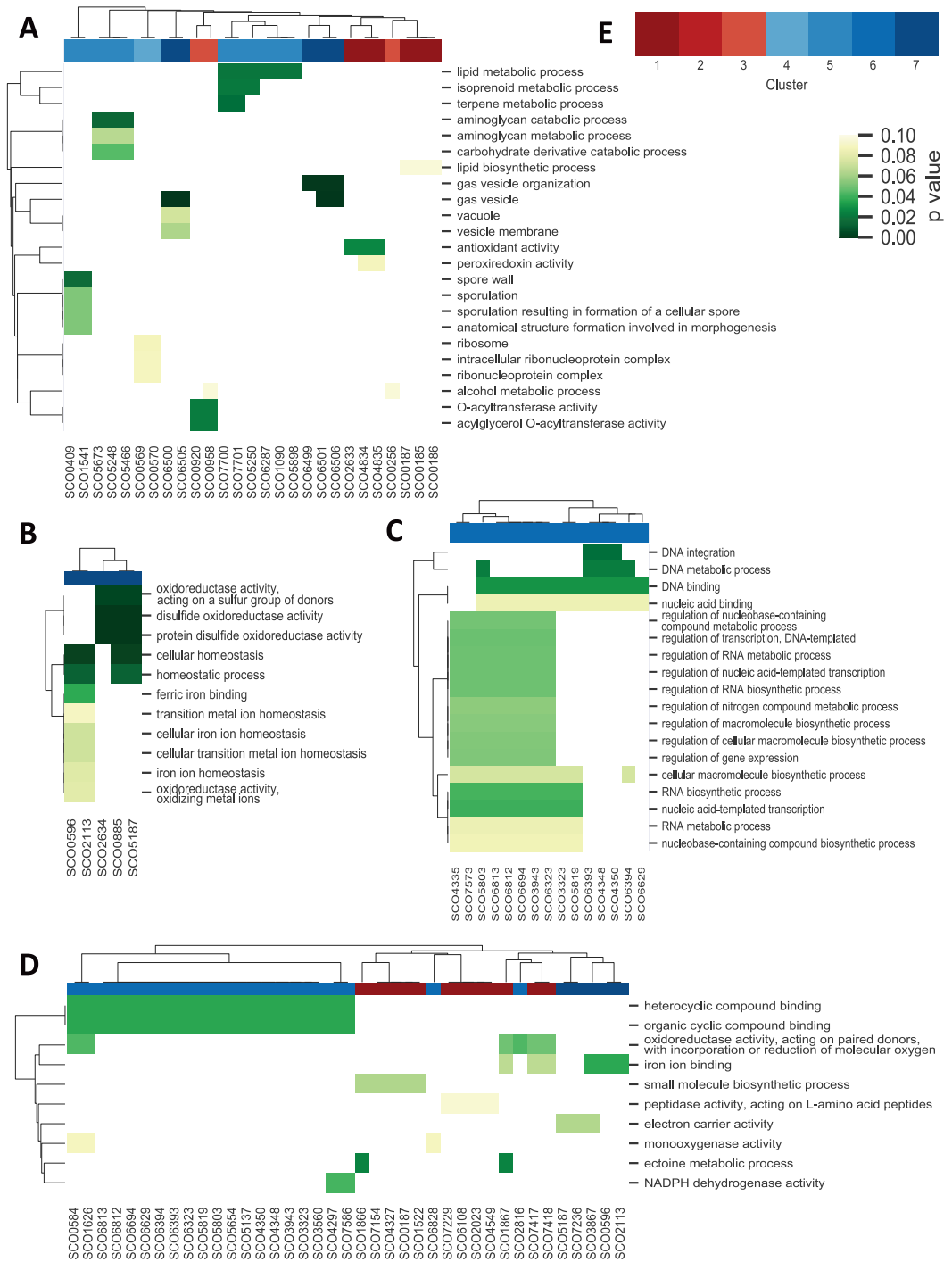


Figure S8: Gene Ontology enrichment analysis of the 7 clusters identified in the 499 differentially expressed genes, categorized by function into four clustered heatmaps. Related to Figure 4, 6 and S7A. Each heatmap shows the p-value for the enrichment of each GO-process. A) Genes related to reactive oxygen species, the ribosome or development process and cell wall formation. B) Oxireductase and iron / metal ion homeostasis. C) Regulation, biosynthesis and metabolism related to RNA and DNA. D) All other GO-annotations. E) This color palette is the legend for the column colors on top of each heatmap which displays which of the seven clusters each gene belongs to. The red palette covers cluster 1-3 (upregulated in M1152), while the blue palette covers cluster 4-7 (upregulated in M145). Note that no GO-processes were enriched for the genes in cluster 2.

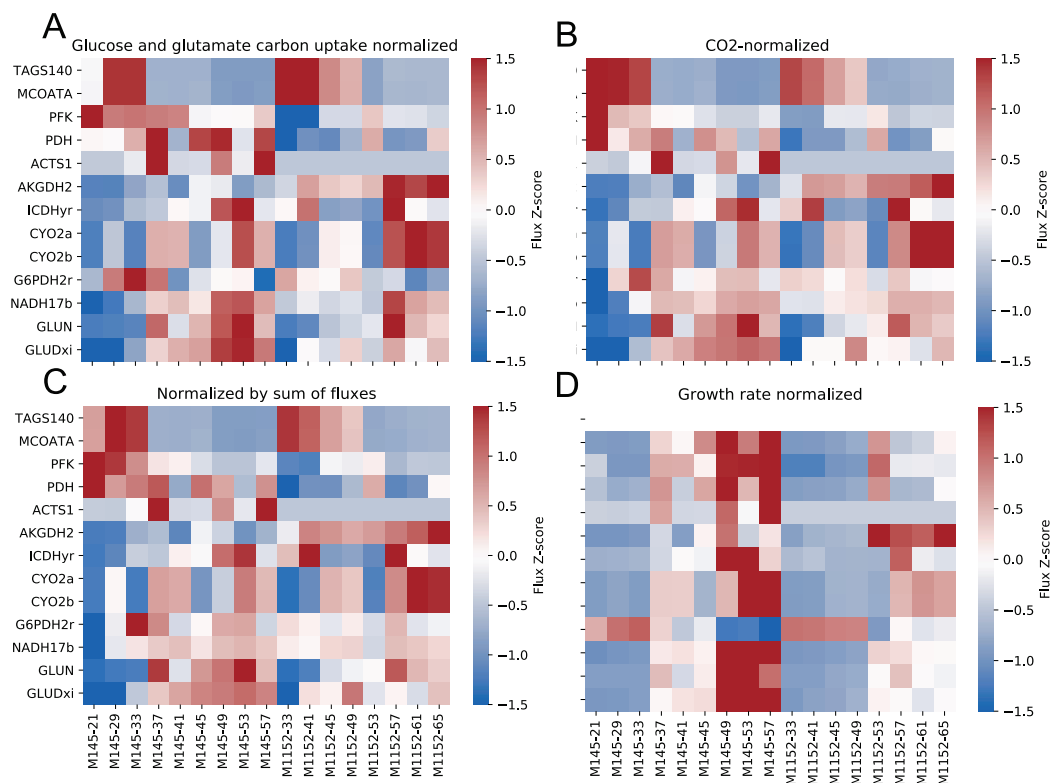


Figure S9: Comparison of normalization methods of randomly sampled fluxes, related to Figure 2D and 3D. Heatmap showing mean flux values normalized by A) total carbon uptake from glucose and glutamate, B) CO₂ production, C) sum of all fluxes and D) growth rate. Because the mean flux values in these reactions are different by several orders of magnitude, we display the data as standardized values (for each reaction).

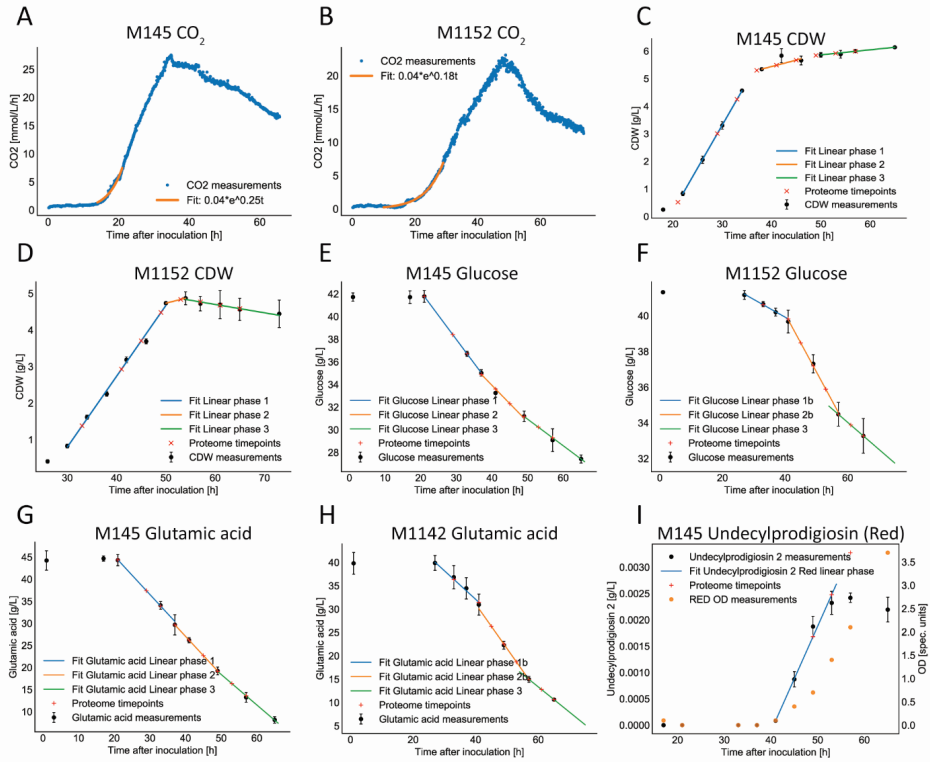


Figure S10: Estimation of rates for M145 and M1152 from cultivation data, related to Transparent methods, Table S1 and S2. A and B) Exponential fit of the CO₂ data to the exponential growth phase of M145 and M1152, respectively. C and D) Piecewise linear fit to estimate growth rates from the CDW measurements of M145 and M1152, respectively. E and F) Piecewise linear fit of glucose concentration in the cultivations of M145 and M1152, respectively. G and H) Piecewise linear fit of glutamic acid concentration in the cultivations of M145 and M1152, respectively. I) Estimated production rate of undecylprodigiosin (Red) in M145.

Table S1: Estimated cell dry weight (CDW) and growth, uptake and secretion rates for M145 at the timepoints of the proteome samples, related to Figure 2D. The unit is mmol/g DW/h for the uptake / secretion rates.

TAI	Estimated CDW [g/L]	Growth rate [h ⁻¹]	Glucose	Glutamic acid	RED	Germicidin-A	Germicidin-B
21	0.517	0.246*	-4.528 [§]	-11.462 [§]	0	0	0
29	3.007	0.103	-0.779	-1.973	0	0	0
33	4.251	0.073	-0.551	-1.395	0	0	0
37	5.301	0.009	-0.338	-1.116	9.60E-05	5.70E-05	8.00E-05
41	5.487	0.008	-0.327	-1.078	9.20E-05	5.50E-05	7.80E-05
45	5.672	0.008	-0.316	-1.043	8.90E-05	5.40E-05	7.50E-05
49	5.845	0.003	-0.223	-0.803	8.70E-05	5.20E-05	7.30E-05
53	5.918	0.003	-0.220	-0.793	8.60E-05	5.10E-05	7.20E-05
57	5.991	0.003	-0.217	-0.784	8.50E-05	5.10E-05	7.10E-05

**This is the maximal growth rate predicted from the exponential fit of the CO₂ curve. The estimated rate from the linear fit of the CDW was unrealistically high.*

§The values for the glucose and glutamate uptake rates are probably too high.

Table S2: Estimated cell dry weight (CDW) and growth, uptake and secretion rates for M145 at the timepoints of the proteome samples, related to Figure 3D. The unit is mmol/g DW/h for the uptake / secretion rates.

TAI	Estimated CDW [g/L]	Growth rate [h ⁻¹]	Glucose	Glutamic acid	RED	Germicidin-A	Germicidin-B
33	1.379	0.140	-0.399	-3.017	0	0	0
41	2.929	0.066	-0.188	-1.421	0	0	0
45	3.704	0.052	-0.394	-1.415	0	0	0
49	4.478	0.043	-0.382	-1.374	0	0	0
53	4.835	0.007	-0.372	-1.336	0	3.00E-06	1.56E-05
57	4.767	-0.005	-0.177	-0.772	0	3.10E-06	1.58E-05
61	4.676	-0.005	-0.180	-0.787	0	3.10E-06	1.61E-05
65	4.585	-0.005	-0.184	-0.803	0	3.20E-06	1.64E-05

Transparent methods

Sco-GEM consensus model reconstruction and development

Sco-GEM, the community consensus model for *Streptomyces coelicolor* is developed, maintained, hosted and publicly available on GitHub (<https://github.com/SysBioChalmers/Sco-GEM>). When we refer to files in the following sections, we use the file names and relative to the main folder in this GitHub repository. By hosting the model on GitHub, we make the reconstruction transparent, the data accessible, provide a structure framework for further development by the community. To this end we also created a channel on Gitter dedicated to Sco-GEM questions and discussions (<https://gitter.im/SysBioChalmers/Sco-GEM>). The model repository was created using memote (Lieven et al., 2018) and we use a [GitFlow structure](#) with two main branches, the *devel* branch contains the most recent changes while the *master* branch contains the stable releases. All new features or bug fixes are performed in separate branches that are incorporated into the *devel* branch through *pull requests*. Semantics for branch names and commit messages are described in *CONTRIBUTING.rst*. The main script language for the model reconstruction is python (version > 3.6), with the exception being the *feat/ecModel* branch with the development of the enzyme-constrained model (EcSco-GEM) where Matlab (version > 7.3) is used.

In terms of folder structure data files, scripts and model files are stored in *ComplementaryData*, *ComplementaryScripts*, and *ModelFiles*, respectively. In the main folder we find the following files:

- *.gitignore*: File which describes file formats automatically ignored by git
- *.gitconfig*: Git config file
- *.gitmodules*: List of linked submodules
- *CONTRIBUTING.rst*: Guidelines describing how to contribute
- *README.md*: General information about the repository
- *HISTORY.rst*: History of model version releases
- *LICENSE.md*: License information
- *memote.ini*: File created by memote (Lieven et al., 2018)
- *requirements.txt*: List of python-packages required to run the model reconstruction

- `.travis.yml`: Config file for automatization of memote with Travis (<https://travis-ci.org/>)

Sco-GEM can be reconstructed at any time using the python script `ComplementaryScripts/reconstruct_scoGEM.py`. Each task of the reconstruction process is performed in a separate script and associated with an issue on GitHub (**Data Set S1, Tab 1**). The details of each task are described in the following paragraphs.

Curate identified issues in iKS1317

We used iKS1317 (Kumelj et al., 2019) as the starting point for the reconstruction of Sco-GEM. Since the publication of iKS1317, several issues had been identified and these were curated as the initial step in the reconstruction pipeline. The curations include correcting the mass and charge balance of the reactions NOR_syn, OAADC, SEPHCHCS and DIOP5OR, and correcting the ec-code, KEGG annotation and gene association for the reactions 3OXCOAT, MMSYNB, PGMT, PPM, ME1, GLUDyi, GLUSx, GLUSy and GLUN.

Curate and add reactions from Sco4

The Sco4 GEM of *S. coelicolor* (Wang et al., 2018) contained additional reactions that we wanted to include in Sco-GEM. However, prior to adding content from Sco4 we curated issues that had been identified since publication. Eleven reactions were found to be either duplicated or wrong in Sco4, and these were removed: RXN0-5224, METHYLGLUTACONYL-COA-HYDRATASE-RXN, GLU6PDEHYDROG-RXN, RXN-15856, 1.14.13.84-RXN_NADPH, R03998, R03999, R09692_NADPH, RXN-9930, 1.17.1.1-RXN_NADH, R09692_NADH. We additionally updated the gene annotations of the following reactions: RMPA, ABTDG, PROD2, THRPDC, ADCL, OXPTNDH, GLNTRS, CU2abc, CBlabc, CBL1abc, GSnt2, INSt2 and PDH.

To enable addition of reactions from Sco4 (Wang et al., 2018) to Sco-GEM we mapped reactions added during the Sco4 development to reactions present in iKS1317 (Kumelj et al., 2019). This mapping was performed semi-automatically: automatic mapping using KEGG and BioCyc annotations followed by manual curation. In total, 394 new reactions and 404 new metabolites were added from Sco4 to Sco-GEM (**Data Set S1, Tab 5 and 6**). Most of the reactions and metabolites added from Sco4 had IDs from the MetaCyc database (Caspi et al.,

2014), containing characters such as dash or parentheses not properly handled by the SBML parser in COBRAPy (Ebrahim et al., 2013). Thus, the ID of all reactions and metabolites added from Sco4 were changed to the correct BiGG ID if possible, otherwise a new ID was created according to the guidelines given in BiGG (King et al., 2016). KEGG (Kanehisa, 2000) and MetaNetX (Moretti et al., 2016) identifiers were included as annotations when possible. Full lists of the IDs and annotations given to reactions and metabolites added from Sco4 are found in the GitHub repository folder *ComplementaryData/curation* as *added_sco4_reactions.csv* and *added_sco4_metabolites.csv*, respectively.

Add gene annotations, reactions and metabolites to Sco-GEM from iAA1259

Based on supplementary files 4 and 5 from iAA1259 (Amara et al., 2018) which list the reactions and metabolites added in iAA1259, we identified 44 reactions and 31 metabolites present in neither Sco4 or iKS1317 (**Data Set S1, Tab 7 and 8**). These 44 reactions were added from iAA1259 and were mainly related to coelimycin biosynthesis, xylan and cellulose degradation and butyrolactones pathway. We further incorporated the modification of 27 reactions curated in iAA1259, associated with oxidative phosphorylation, futasoline pathway or chitin degradation (**Data Set S1, Tab 9**). These curations mainly updated gene-reaction rules but also updated reaction bounds and deletion of two reactions (CFL and DHFUTALS). Finally, we incorporated the biomass-function which was updated in iAA1259.

Change direction of reactions that were backwards irreversible

The pipeline for reconstruction of the enzyme-constrained model required all reactions to be either reversible or forward irreversible (i.e. reactions with bounds $(-1000, 0)$ are not allowed). Therefore, all backward irreversible reactions were rewritten (substrates were changed to products and *vice versa*) so they could be represented as forward irreversible.

Fix missing / wrongly annotated reactions and metabolites

We identified several minor issues related to reaction and metabolite IDs or annotations. These may come from the current or previous model reconstruction efforts. These issues include:

- Misspelled IDs or annotations
- Empty annotations in SBML file
- Wrong BioCyc annotations for metabolites and reactions in the germicidin pathway

- Update all MetaNetX annotations
- Exchange reactions given BioCyc annotations
- Fix chebi annotations so they comply with the MIRIAM identifiers
- Mixed up IDs for actACPmmy and malACPmmy

Create pseudo-metabolites for NADH/NADPH and NAD⁺/NADP⁺ to use in reaction where the redox cofactor is not known

For some redox reactions added from Sco4, it was not sure if NADH/NAD⁺ or NADPH/NADP⁺ was the participating cofactor pair. In this case, both possibilities were included in Sco4. However, to avoid duplicated reactions and make it explicit that the cofactor is unknown we changed these reactions to use pseudo-metabolites (acceptor_c and donor_c) as the cofactor pair. We then also included pseudo-reactions which converts NADH/NADPH and NAD⁺/NADP⁺ to donor_c and acceptor_c, respectively [pseudo-reaction IDs: PSEUDO_DONOR_NADH; PSEUDO_DONOR_NADPH; PSEUDO_ACCEPTOR_NAD; PSEUDO_ACCEPTOR_NADP]. In total 17 enzymatic reactions use these pseudo-metabolites as cofactor pair: 3OCHOCDH; OXCOADH; 4DPCDH; 4HYDPRO; 4NITROB; AHLGAL; AHOPS; CADHX; DDALLO; DPCOX; GDP64HRD; HDAPMO; PHYFLUDS; HYTDES; SORBDH; ZCARDS; ZCAROTDH2.

Add SBO terms to genes, reactions and metabolites

SBO (Systems Biology Ontology) (Courtot et al., 2011) terms were included as annotations of reactions, genes and metabolites according to **Data Set S1, Tab 10**.

Update the biomass reaction

In iAA1259, the biomass reaction was curated in respect to 2-demethylmenaquinol and menaquinol, however, this resulted in a biomass reaction that combined described more than 1 g per gDCW. In addition, the biomass reaction of all *S. coelicolor* models have described small molecule and protein co-factors/prosthetic groups as components, where their abundance was arbitrarily set to complement the remaining biomass components to reach 1 g per gDCW. This is likely a gross overestimation for many of these molecules, and this proved problematic for initial simulations with the enzyme constrained model. In contrast to enzymes of central carbon metabolism, enzymes involved in biosynthesis of such co-factors and prosthetic groups

have typically lower efficiency, such that large fractions of the protein allocation would have to be devoted to these pathways if the abundances are overestimated.

The availability of proteomics data has allowed us to give more reasonable estimates of abundance of protein-linked cofactors and prosthetic groups. The new biomass reaction was estimated through the following steps:

1. By querying UniProt, a list of prosthetic groups per protein were collated (*ComplementaryData/biomass/prosthetic_groups_uniProt.txt*) and further processed (*ComplementaryScripts/ecModel/prostheticGroups.m*) as detailed below.
2. If *metal* was specified as cofactor, the abundance was split over cobalt²⁺, copper²⁺, iron²⁺, zinc²⁺, nickel²⁺, calcium²⁺, potassium⁺, magnesium²⁺ and manganese²⁺.
3. Dipyrromethane is generated by the enzyme itself from its substrate and is therefore not further considered.
4. From the M145 and M1152 cultivation data, quantitative proteomics was estimated as detailed below.
5. Cofactor abundances were estimated by combining the estimated protein levels and the protein cofactor annotation (available at *ComplementaryData/biomass/prosthetic_groups_mets.txt*)
6. To simplify fitting of biomass components, the full biomass reaction was split into the pseudometabolites *lipid*, *dna*, *rna*, *protein*, *carbohydrate*, *cell_wall* and *misc*, with the latter containing the cofactors (*ComplementaryData/biomass/standard_biomass.txt*).
7. After updating the abundances of the cofactors, the remaining *misc* metabolites were refitted to ensure that the total biomass adds up to 1 g per gDCW.

The updated composition (*ComplementaryData/biomass/biomass_scaled.txt*) was subsequently used to modify the model stoichiometry (*fix_biomass.py*). A comparison of the updated biomass reaction and the biomass reaction in iAA1259 is presented in **Data Set S1, Tab 2**.

Model reversibility

By using the python-API (<https://gitlab.com/elad.noor/equilibrador-api>) of eEquilibrator (Flamholz et al., 2012) we calculated the change in Gibbs free energy for 770 reactions (**Data**

Set S1, Tab 3). eQuilibrator can only calculate the change in Gibbs free energy for intracellular reactions (i.e. not transport and exchange reactions) where all metabolites are mapped to KEGG (Kanehisa, 2000; Kanehisa et al., 2019). The calculations are based on the component contribution method (Noor et al., 2013). The change in Gibbs free energy was calculated at standard conditions (25 °C, 1 bar), pH7 and 1mM concentration of reactants, denoted $\Delta_r G^m$ in eQuilibrator. We then applied a threshold of -30 kJ/mol to define a reaction as irreversible (Bar-Even et al., 2012; Feist et al., 2007), and compared the calculated reversibility with the reversibility of these reactions in the model prior to curation. We found that the reversibility was equal for 56.9% (438 / 770) of the reactions (**Figure 1E**). The majority of differences were reactions that were irreversible in the model but classified as reversible using the calculated values for the change in Gibbs free energy (35%; 273/770; **Figure 1E**).

Using the set of growth data and knockout data, we evaluated the effect of the suggested changes in reaction reversibility: by randomly applying these changes to 10 reactions at the time, we identified 13 single, 22 pairs and 13 triplets of reactions (**consisting of 55 unique reactions**) that reduced model accuracy when the reversibility was changed based on the change in Gibbs free energy (**Data Set S1, Tab 11**). Then we used the data set of growth and gene knockout phenotypes (Kumelj et al., 2019) to identify another 6 reactions that caused erroneous predictions if the reversibility were changed (PROD2, ARGSS, OCT, URIK1, URIK2, and UPPRT). These 61 reactions were discarded from having the reversibility changed, resulting in a total of 271 reactions with changed reversibility.

Energetic cofactors, including ATP, NADPH, NADH, FAD and any quinone, were involved in 284 of the 770 reactions for which the change in Gibbs free energy was calculated. Of the 114 reactions involving ATP, 82 reactions had an estimated change in Gibbs free energy between ± 30 kJ/mol, indicating that the reactions were reversible. Because one assumes that ATP-driven reactions in general are irreversible (Thiele and Palsson, 2010), the reversibility of these 82 reactions were manually curated (**Data Set S1, Tab 12**). For the 7 quinone-associated reactions for which the change in Gibbs free energy was calculated (CYTBD2, NADH17b, NADH10b, MBCOA2, G3PD5, PDH3, NADH2r) all were defined as irreversible as previously suggested (Thiele and Palsson, 2010). The reversibility of reactions involving any of the other energetic cofactors were treated as any other reaction as previously described.

Analysis and annotation of transport reactions

Gene annotations, substrate and transport class information were mostly extracted from Transport DB 2.0 (Elbourne et al., 2017) and TCDB (Saier et al., 2016). Then, transport proteins were extracted from IUBMB-approved Transporter Classification (TC) System and categorized into 9 main classes (**Figure 1F**): 1) ABC transporter; 2) PTS transporter; 3) Proton symporter; 4) Sodium symporter; 5) Other symporter; 6) Proton antiporter; 7) Other antiporter; 8) Facilitated diffusion; 9) Simple diffusion. For those transport proteins with an ambiguous substrate annotation in TCDB, the specific substrate annotation was obtained by extracting annotations from KEGG (Kanehisa, 2000; Kanehisa et al., 2019), UniProt (The UniProt Consortium, 2019) or through BLAST homology search (NCBI Resource Coordinators, 2017) using a similarity threshold of 90% (**Data Set S1, Tab 4**).

Subsystem annotations

We leveraged the KEGG and BioCyc annotations of each individual reaction to extract a draft subsystem and pathway annotation for each reaction. For KEGG, this was achieved by using the python module BioServices (Cokelaer et al., 2013) while we used PythonCyc (<https://github.com/latendre/PythonCyc>) and PathwayTools (Karp et al., 2016) to extract pathway annotations from BioCyc (Karp et al., 2019).

The draft annotations were then curated, and each reaction was annotated to one out of 15 subsystems. When no or multiple annotations were extracted from the databases we used adjacent reactions in the metabolic network to infer the single, most correct annotation. These 15 subsystem categories are based on the categories of the KEGG Pathway Maps for metabolism (<https://www.genome.jp/kegg/pathway.html>) but we have included three additional categories to cover all aspects of the model: Biomass and maintenance functions, Membrane Transport, and Exchange (**Figure S1**).

We also annotated 1964 of the 2552 reactions to one out of 128 different pathways. The remaining 588 are mostly transport and exchange reactions, or possibly reactions not fitting into any of these pathways.

Export model file with alphabetical ordering

An import feature with GitHub is the ability to easily see changes in text files after every commit. However, COBRAPy (Ebrahim et al., 2013) doesn't sort the list of reactions, metabolites and genes before the SBML-file is written and this makes it look like there was a

lot of changes even when the model is unchanged. Thus, we now sort these lists before writing to file. The export function also stores the model-file in the YAML format which is more readable than SBML (XML). Finally, the export function creates the *requirements.txt* file which holds information about all non-standard python modules necessary to run the model reconstruction.

Development of enzymatically constrained (EcSco-GEM) model

An enzyme-constrained version of the Sco-GEM model (denoted EcSco-GEM) was generated using GECKO (Sánchez et al., 2017). The GECKO method enhances an existing GEM by explicitly constraining the maximum flux through each reaction by the maximum capacity of the corresponding enzyme, given by the product of the enzyme abundance and catalytic coefficient. Both reversible reactions and reactions catalysed by isoenzymes (redundant genes) are handled automatically by the GECKO method by splitting each occurrence into individual reactions. The Sco-GEM v1.1 model was modified using GECKO version 1.3.4. Kinetic data, in the form of k_{cat} values (s^{-1}), were automatically collected from BRENDA (Jeske et al., 2019). If BRENDA did not report a k_{cat} value for an enzyme, GECKO searched for alternative k_{cat} values by reducing specificity, on the level of substrate, enzymatic activity (EC number) and organism.

A total of 4753 k_{cat} values were matched, including separate values for forward and backward direction for reversible reactions, of which:

- 53 were matched with organism (*S. coelicolor*) and correct substrate
- 1541 were matched with closest organism and correct substrate
- 236 were matched with organism (*S. coelicolor*) and any substrate
- 15 were matched with organism (*S. coelicolor*) and any substrate, reported specific activities instead of k_{cat} (corrected in the model for molecular weight of the enzyme)
- 2586 were matched with closest organism and any substrate
- 322 were matched with any organism and any substrate, reported specific activities instead of k_{cat} (corrected in the model for molecular weight of the enzyme)

The algorithm first looped through these criteria above, with the full EC code. If no match

could be found, wildcards were added (e.g. EC2.3.4.- instead of EC2.3.4.5), followed by going through the list of criteria above. The statistics there is:

- 4178 were matched without any wildcards (full EC code)
- 544 were matched after adding one wildcard
- 21 were matched after adding two wildcards (e.g. EC2.3.-.-)
- 10 were matched after adding three wildcards
- 0 were matched after adding four wildcards

Using the initial set of BRENDA-suggested k_{cat} values, the model was evaluated to support simulation of experimentally measured growth rates. During this testing the NAD(H)/NAD(P)H pseudo-reactions were blocked to avoid infeasible loops.

The following k_{cat} values were identified as growth limiting resulting in the stated manual curations:

- Chorismate synthase (CHORS; EC4.2.3.5; SCO1496; Q9KXQ4)
Sco-GEM uses 5-O-(1-Carboxyvinyl)-3-phosphoshikimate as name of the main substrate, while BRENDA uses its synonym 5-enolpyruvylshikimate 3-phosphate. This prevented automatically finding the substrate. Hence, the k_{cat} was manually changed to 0.87 s^{-1} , as measured from *N. crassa* (Rauch et al., 2008).
- Phosphoribosylformylglycinamide synthase (PRFGS; EC6.3.5.3; SCO4077 and SCO4078 and SCO4079; Q9RKK5 and Q9RKK6 and Q9RKK7)
 k_{cat} suggested by BRENDA used NH_4^+ as substrate, instead of glutamine. Specific activity using glutamine is provided for *E. coli*: $2.15 \text{ } \mu\text{mol}/\text{min}/\text{mg}$ protein (Schendel et al., 1989). Assuming molecular weight of 141 kDa, this translates to $k_{cat} = 5.05 \text{ s}^{-1}$.
- Methylmalonate-semialdehyde dehydrogenase (malonic semialdehyde) (MMSAD3; EC1.2.1.27; SCO2726; Q9L1J1)
 k_{cat} suggested by BRENDA is from archaea, instead use k_{cat} value of 2.2 s^{-1} from *B. subtilis* (Talfournier et al., 2011).

- Phosphoribosyl-ATP pyrophosphatase (PRATPP; EC3.6.1.31; SCO1439; Q9EWK0)
 k_{cat} suggested by BRENDA was calculated from specific activity in *Salmonella enterica*, but the reported value was measured in cell extract, not from purified enzyme. Instead, use specific activity from *S. cerevisiae*: 332 $\mu\text{mol}/\text{min}/\text{mg}$ protein (Keesey et al., 1979). Assuming molecular weight of 95 kDa, this translates to a k_{cat} of 526 s^{-1} .
- Glyceraldehyde 3-phosphate dehydrogenase (Q9Z518/EC1.2.1.12) - assigned k_{cat} from *Corynebacterium glutamicum* was highly growth limiting. Instead use specific activity measured of pentalenolactone sensitive gapdh in *Streptomyces arenae*: 112 $\mu\text{mol}/\text{min}/\text{mg}$ protein (Maurer et al., 1983).

Then, separate models were created for each strain (the gene clusters for actinorhodin, undecylprodigiosin, CDA and coelimycin P1 were removed to create M1152) and for each time point by using estimated growth, uptake rates of glutamate and glucose, secretion rates of undecylprodigiosin, germicidin A and B and proteome measurements. The estimated growth, uptake and secretion rates were estimated from raw measurements across three biological replicates (details provided in the last section). These time point specific models (9 time points for M145, 8 time points for M1152) were used to analyse the activity in individual metabolic pathways through random sampling (Bordel et al., 2010). We also created one EcSco-GEM model for each strain with a global constraint on the protein usage instead of specific protein usage, which were used for model quality control.

[Continuous integration and quality control with memote](#)

Validation and quality assessment of Sco-GEM is carried out using the test-suite in memote (Lieven et al., 2018). Memote provides by default a large range of tests, which we have used to identify issues and possible improvements. The test suite reports descriptive model statistics such as the number of genes, reactions and metabolites, and also checks the presence of SBO terms and annotations, the charge and mass balance of all reactions, the network topology and find energy-generating cycles (Fritzemeier et al., 2017). Additionally, we incorporated custom tests into the memote test-suite to automatically compare predicted phenotypes with experimental data in different growth media and for different knockout mutants. In addition to the classical binary classifiers accuracy, sensitivity and specificity we also report the Matthews correlation coefficient which is considered to be more reliable when

the number of elements in each classification category is skewed (Chicco and Jurman, 2020). The Matthews correlation coefficient (*MCC*) is calculated from the true positive (*TP*), false positive (*FP*), true negative (*TN*) and false negative (*FN*) values as $MCC = \frac{TP \cdot TN - FP \cdot FN}{\sqrt{(TP+FP) \cdot (TP+FN) \cdot (TN+FP) \cdot (TN+FN)}}$. The experimental growth and knockout data are extracted from (Kumelj et al., 2019). As a separate evaluation, we applied another method for identifying internal and unrealistic energy-generating cycles (Noor, 2018), and no such cycles were found in Sco-GEM.

The simplest use of memote is generating snapshot reports showing the current state of the model. However, by integrating Travis CI [<https://travis-ci.com/>] into the gitHub repository, memote can be used to create a continuous report displaying how each commit affects the model quality. Memote version 0.9.12 was used in this work, and the memote snapshot report for Sco-GEM is given in the **Supplemental Information**.

Random sampling, normalization and pathway analysis

Because of the huge number of reactions in the EcSco-GEM, it is challenging to sample the solution space appropriately: we have chosen to use the method provided in the Raven Toolbox 2 (Bordel et al., 2010; Wang et al., 2018), which samples the vertices of the solution space. The drawback of this method is that it will not result in a uniform sampling of the solution space. However, it is more likely to span the entire solution space and also not prone to get stuck in extremely narrow parts of the solution space, which may happen with variants of the hit-and-run algorithm (Haraldsdóttir et al., 2017; Kaufman and Smith, 1998; Megchelenbrink et al., 2014). For each of the time points for each strain (17 different conditions in total) we constrained exchange reactions between 99% and 101% of the measured rates and generated 5000 random flux distributions with Gurobi as the solver. The reactions catalysed by isoenzymes were combined into the set of reactions in Sco-GEM and the reactions providing protein for each reaction. The mean of the 5000 flux distributions for each metabolic reaction was used in the following analysis.

Finally, for each of the 17 conditions, the mean fluxes were normalized by the CO₂ production rate. Then, the normalized mean fluxes were summarized for each metabolic pathway by using the curated pathway annotations, and we consider this a measure of the metabolic activity in

each pathway. To ease visual interpretation of this data we used the function `clustermap` with default parameters in Seaborn version 0.9.0 (Michael Waskom et al., 2018) (which uses Scipy v1.3.1 (Virtanen et al., 2020)) to perform hierarchical clustering of pathways based on the metabolic activity in M145 (**Figure 2D**). We then kept this order in **Figure 3D** to enable strain comparison.

Since glucose and glutamate uptake rates, as well as growth rates were significantly different in the two strains and at different time points, normalization of the data was necessary to compare flux distributions. We tested various proxies as indicators of overall metabolic activity for normalization, namely CO₂ production; the total carbon uptake from glucose and glutamate; growth rate and mean flux value. As golden standard, we compared the fluxes through individual reactions that are well documented to change in M145 in response to the phosphate depletion (**Figure S9**). Normalization based on CO₂ production was tested and gave similar results than the data normalized on total carbon uptake from glucose and glutamate (**Figure S9A and S9B**). The data normalized by the sum of fluxes showed similar patterns as those achieved by glucose/glutamate and CO₂-normalized data but was noisier (**Figure S9C**). Considering the huge differences in growth rate, the growth-normalized data masked any other flux patterns (**Figure S9D**). The fact that different normalizations provided similar differences in metabolic fluxes proved that the inferred changes in metabolism were not artefacts of the normalization method but represent true metabolic activity of each strain.

[Strains, cultivation conditions, sampling procedures, and analyses of media components and secondary metabolites](#)

Experiments were performed using strain M145 of *S. coelicolor* A3(2) and its derivatives M1146 and M1152. The latter two are lacking the 4 major BGCs for actinorhodin (Act), undecylprodigiosin (Red), coelimycin P1 (Cpk), and calcium-dependent antibiotic (CDA), while M1152 is also carrying the pleiotropic, previously described antibiotic production enhancing mutation *rpoB* [S433L] (Gomez-Escribano and Bibb, 2011; Hu et al., 2002). All strains were kindly provided by Mervyn Bibb at John-Innes-Centre, Norwich, UK.

Triplicate cultivations of the strains were performed based on germinated spore inoculum on 1.8 L phosphate-limited medium SSBM-P, applying all routines of the optimized submerged batch fermentation strategy for *S. coelicolor* established and described before (Wentzel et al., 2012). All media were based on ion-free water, and all chemicals used were of analytical grade.

In brief, spore batches of M145, M1146 and M1152 were generated by cultivation on soy flour-mannitol (SFM) agar plates (Kieser et al., 2000), harvesting by scraping off spores and suspension in 20% (v/v) glycerol, and storage in aliquots at $-80\text{ }^{\circ}\text{C}$. 10^9 CFU of spores of each strain were germinated for 5 hours at $30\text{ }^{\circ}\text{C}$ and 250 rpm in 250 mL baffled shake-flasks with 2 g of 3 mm glass beads and 50 mL 2x YT medium (Claessen et al., 2003). The germinated spores were harvested by centrifugation ($3200\text{ } \times\text{ } g$, $15\text{ }^{\circ}\text{C}$, 5 min) and re-suspended in 5 mL ion-free water. An even dispersion of the germinated spores was achieved by vortex mixing (30 s), ensuring comparable inocula among biological replicas. Each bioreactor (1.8 liter starting volume culture medium in a 3-liter Applikon stirred tank reactor) was inoculated with 4.5 mL germinated spore suspension (corresponding to 9×10^8 CFU). Phosphate-limited medium SSBM-P (Nieselt et al., 2010) consisted of Na-glutamate, 55.2 g/L; D-glucose, 40 g/L; MgSO_4 , 2.0 mM; phosphate, 4.6 mM; supplemented minimal medium trace element solution SMM-TE (Claessen et al., 2003), 8 mL/L and TMS1, 5.6 mL/L. TMS1 consisted of $\text{FeSO}_4 \times 7\text{ H}_2\text{O}$, 5 g/L; $\text{CuSO}_4 \times 5\text{ H}_2\text{O}$, 390 mg/L; $\text{ZnSO}_4 \times 7\text{ H}_2\text{O}$, 440 mg/L; $\text{MnSO}_4 \times \text{H}_2\text{O}$, 150 mg/L; $\text{Na}_2\text{MoO}_4 \times 2\text{ H}_2\text{O}$, 10 mg/L; $\text{CoCl}_2 \times 6\text{ H}_2\text{O}$, 20 mg/L, and HCl, 50 mL/L. Clerol FBA 622 fermentation defoamer (Diamond Shamrock Scandinavia) was added to the growth medium before inoculation. Throughout fermentations, pH 7.0 was maintained constant by automatic addition of 2 M HCl. Dissolved oxygen levels were maintained at a minimum of 50% by automatic adjustment of the stirrer speed (minimal agitation 325 rpm). The aeration rate was constant 0.5 L/(L \times min) sterile air. Dissolved oxygen, agitation speed and carbon dioxide evolution rate were measured and logged on-line, while samples for the determination of cell dry weight, levels of growth medium components and secondary metabolites concentrations, as well as for transcriptome and proteome analysis were withdrawn throughout the fermentation trials as indicated in **Figure 2B**. For transcriptome analysis, 3×4 ml culture sample were applied in parallel onto three $0.45\text{ }\mu\text{m}$ nitrocellulose filters (Millipore) connected to vacuum. The biomass on each filter was immediately washed twice with 4 ml double-autoclaved ion-free water pre-heated to $30\text{ }^{\circ}\text{C}$, before the filters were collected in a 50 ml plastic tube, frozen in liquid nitrogen and stored at $-80\text{ }^{\circ}\text{C}$ until RNA isolation. For proteome analysis, 5 ml samples were taken and centrifuged ($3200\text{ } \times\text{ } g$, 5 min, $4\text{ }^{\circ}\text{C}$), and the resulting cell pellets frozen rapidly at $-80\text{ }^{\circ}\text{C}$ until further processing.

Levels of phosphate were measured spectrophotometrically by using the SpectroQuant Phosphate test kit (Merck KGaA, Darmstadt, Germany) following the manufacturer's instructions after downscaling to 96-well plate format. D-glucose and L-glutamate concentrations were determined by LC-MS using suitable standards, and measured concentrations were used to estimate specific uptake and excretion rates. Undecylprodigiosin (Red) levels were determined spectrophotometrically at 530 nm after acidified methanol extraction from the mycelium (Bystrykh et al., 1996). To determine relative amounts of actinorhodins (determined as total blue pigments, TBP), cell culture samples were treated with KOH (final concentration 1 M) and centrifuged, and the absorbance of the supernatants at 640 nm was determined (Bystrykh et al., 1996). Quantification of germicidin A and B was performed using targeted LC-MS analytics.

Proteomics

Sample preparation and NanoUPLC-MS analysis

Quantitative proteomics were performed using pipeline previously described (Gubbens et al., 2012). Mycelium pellets for proteome analysis were thawed and resuspended in the remaining liquid. 50 μ L re-suspended mycelium was withdrawn and pelleted by centrifugation. 100 μ L lysis buffer (4% SDS, 100 mM Tris-HCl pH 7.6, 50 mM EDTA) was added, and samples were sonicated in a water bath sonicator (Biorupter Plus, Diagenode) for 5 cycles of 30 s high power and 30 s off in ice water. Cell debris was pelleted and removed by centrifugation. Total protein was precipitated using the chloroform-methanol method described before (Wessel and Flügge, 1984). The pellet was dried in a vacuum centrifuge before dissolving in 0.1% RapiGest SF surfactant (Waters) at 95 °C. The protein concentration was measured at this stage using BCA method. Protein samples were then reduced by adding 5 mM DTT, followed by alkylation using 21.6 mM iodoacetamide. Then trypsin (recombinant, proteomics grade, Roche) was added at 0.1 μ g per 10 μ g protein. Samples were digested at 37 °C overnight. After digestion, trifluoroacetic acid was added to 0.5% followed by incubation at 37 °C for 30 min and centrifugation to remove MS interfering part of RapiGest SF. Peptide solution containing 8 μ g peptide was then cleaned and desalted using STAGE-Tipping technique (Rappsilber et al., 2007). Final peptide concentration was adjusted to 40 ng/ μ L using sample solution (3% acetonitrile, 0.5% formic acid) for analysis.

200 ng (5 μ L) digested peptide was injected and analysed by reversed-phase liquid chromatography on a nanoAcquity UPLC system (Waters) equipped with HSS-T3 C18 1.8 μ m, 75 μ m X 250 mm column (Waters). A gradient from 1% to 40% acetonitrile in 110 min (ending with a brief regeneration step to 90% for 3 min) was applied. [Glu¹]-fibrinopeptide B was used as lock mass compound and sampled every 30 s. Online MS/MS analysis was done using Synapt G2-Si HDMS mass spectrometer (Waters) with an UDMS^E method set up as described in (Distler et al., 2014).

Data processing and label-free quantification

Raw data from all samples were first analysed using the vendor software ProteinLynx Global SERVER (PLGS) version 3.0.3. Generally, mass spectrum data were generated using an MS^E processing parameter with charge 2 lock mass 785.8426, and default energy thresholds. For protein identification, default workflow parameters except an additional acetyl in N-terminal variable modification were used. Reference protein database was downloaded from GenBank with the accession number NC_003888.3. The resulted dataset was imported to ISOQuant version 1.8 (Distler et al., 2014) for label-free quantification. Default high identification parameters were used in the quantification process. TOP3 result was converted to PPM (protein weight) and send to the modelers and others involved in interpreting the data (**Data Set S3**).

TOP3 quantification was filtered to remove identifications meet these two criteria: 1. identified in lower than 70% of samples of each strain and 2. sum of TOP3 value less than 1×10^5 . Cleaned quantification data was further subjected to DESeq2 package version 1.22.2 (Love et al., 2014) and PCA was conducted after variance stabilizing transformation (vst) of normalized data.

Transcriptomics

RNA extraction and quality control

Bacteria were lysed using RNAprotect Bacteria (Qiagen) and following the manufacturer's instruction. Briefly, filters containing bacteria were incubated with 4 ml of RNAprotect Bacteria reagent. After centrifugation, resulting samples were lysed using 500 μ l of TE buffer (10 mM Tris-Cl, 1 mM EDTA, pH 8.0) containing 15 mg/ml lysozyme using 150-600 μ m diameter glass beads (Sigma) agitated at 30 Hz for 5 minutes in the TissueLyser II (Qiagen).

Total RNA was extracted using RNeasy mini kit (Qiagen) and 700 µl of the resulting lysate complemented with 470 µl of absolute ethanol. RNAase-free DNase set (Qiagen) and centrifugation steps were performed to prevent DNA and ethanol contamination. Elution was performed using 30 µl of RNase-free water and by reloading the eluate on the column to improve the RNA yield. The RNA concentration was measured using Qubit RNA BR Assay Kit (ThermoFisher Scientific), RNA purity was assessed using A260/A280 and A260/A230 ratio using the Nano Drop ND-1000 Spectrophotometer (PEQLAB). RNA Integrity Number was estimated using RNA 6000 Nano Kit (Agilent) and the Bioanalyzer 2100 (Agilent).

Library preparation and sequencing

A total of 1 µg of total RNA was subjected to rRNA depletion using Ribo-Zero rRNA Removal Kit Bacteria (Illumina). The cDNA libraries were constructed using the resulting tRNA and the NEBNext Ultra II Directional RNA Library Prep Kit (NEB). Libraries were sequenced as single-reads (75 bp read length) on an Illumina NextSeq500 platform at a depth of 8–10 million reads each.

RNA-seq data assessment and analysis

Sequencing statistics including the quality per base and adapter content assessment of resulting transcriptome sequencing data were conducted with FastQC v0.11.5 (Andrews, 2016). All reads mappings were performed against the reference strain of *Streptomyces coelicolor* A3(2) (RefSeq ID NC_003888.3). The mappings of all samples were conducted with HISAT2 v2.1.0 (Kim et al., 2015). As parameters, spliced alignment of reads was disabled, and strand-specific information was set to reverse complemented (HISAT2 parameter --no-spliced-alignment and --rna-strandness "R"). The resulting mapping files in SAM format were converted to BAM format using SAMtools v1.6 (Li et al., 2009). Mapping statistics, including strand specificity estimation, percentage of mapped reads and fraction exonic region coverage, were conducted with the RNA-seq module of QualiMap2 v2.2.2-dev (Okonechnikov et al., 2016). Gene counts for all samples were computed with featureCounts v1.6.0 (Liao et al., 2014) based on the annotation of the respective reference genome, where the selected feature type was set to transcript records (featureCounts parameter -t transcript).

Normalization and differential gene expression

Raw count files were imported into Mayday SeaSight (Battke and Nieselt, 2011) for common, time-series-wide normalization. For this, the raw counts of all biological replicates of one strain

across the time-series were log₂-transformed (with pseudocount of +1 for the genes with zero counts) and then quantile-normalized. To make the two normalized time-series data of M154 and M1152 comparable, they were again quantile-normalized against each other. The normalized RNA-seq data are provided in **Data Set S4**.

Differentially expressed genes were identified by ANOVA using Orange (v3.2) and the bioinformatic toolkit (v), with FDR of <0.01 and a minimal fold enrichment >1 for at least one aligned time point. Genes with low expression (log₂ < 5 for both strains and time points) were not considered for further analysis. The differentially expressed genes were subsequently scaled to the expression average and clustered by K-means. Visualization of genes and clusters were performed in python (v3.7) with matplotlib (v3.1.1). For this, the time-series of M145 and M1152 were aligned such that in the visual representation, the expression profiles of the two strains are aligned relative to the time point of phosphate depletion. Both DAVID (Huang et al., 2009a, 2009b) and the string database (Szklarczyk et al., 2019) was used to evaluate the function of each cluster, identifying overrepresentation of function groups based on GO annotation or text mining. Identified differential clusters or regulons were extracted from literature and plotted (**Data Set S2; Figure S8**). When we display the RNA-seq data as heatmaps (**Figure 6 and S3**) the order of genes is determined by hierarchical clustering using methods as previously described for the clustering of pathways based on metabolic activity.

[Estimation of growth, uptake and production rates for *Streptomyces coelicolor* M145 and M1152 from batch fermentation data](#)

The estimated growth, uptake and secretion rates are based on average values of online and offline measurements of batch fermentation from three parallel bioreactors for each strain.

[Growth rate estimation](#)

Both the CDW (cell dry weight) measurements and CO₂ measurements can in principle be used to estimate growth rates, as there should be a linear relationship between the CO₂ concentration and cell mass. The CO₂ concentration is measured online on a high-resolution timescale (5 min) while CDW is measured offline with a four-hour resolution starting from 18 hours after inoculation.

To estimate growth rates, we separated the growth into 5 different phases:

1. Lag phase - immediate phase after inoculation with no / low growth
2. Exponential growth - rapid growth after the initial lag phase

3. First linear growth rate - until phosphate depletion
4. Second linear growth rate – immediate phase after phosphate where there is still growth
5. Third linear growth rate – no or very low growth

From both **Figure 2A and 3A** we find a clear discrepancy between the CO₂ curve and the CDW measurements after phosphate depletion. Thus, despite the lower resolution we decided to use the CDW measurements for the growth rate estimation except for the exponential growth phase.

Exponential growth rate from CO₂

The exponential growth rate was estimated by fitting an exponential curve on the form

$$X(t) = X_0 e^{\mu t}$$

to the selected region of the CO₂ measurements (**Figure S10A and S10B**), leading to an estimated growth rate of $0.25 \text{ h}^{-1} \pm 0.06 \text{ h}^{-1}$ and $0.18 \text{ h}^{-1} \pm 0.02 \text{ h}^{-1}$ for M145 and M1152, respectively. The uncertainty is estimated in a heuristic approach by observing the minimum and maximum values observed when changing the boundaries for the fitted function.

Linear growth rates from CDW

The growth rate is estimated by fitting linear slopes to the three different linear phases of growth (**Figure S10C and S10D**). The specific growth rate is then calculated using the following equation

$$\frac{dX}{dt} = \mu X \rightarrow \mu = \frac{1}{X} \cdot \frac{dX}{dt}$$

where μ is the growth rate, X the CDW and $\frac{dX}{dt}$ is the slope of the linear fit. Because the inverse of the CDW the rates can become very large when the cell mass is low, but we use the estimated growth rate in the exponential phase as an upper bound. Predicted growth rates and CDW estimates at the timepoints for the proteome samples are given for M145 and M1152 in **Table S1** and **Table S2**, respectively.

Uptake rates of glucose and glutamic acid

The uptake rates for glucose and glutamate were also fitted using a piecewise linear function (**Figure S10E-H**). Using the same time intervals as for the CDW estimates gave a very poor fit

for M1152 and we therefore decided to use different time intervals. From the fitted slopes we estimated the uptake rates using the equation given below:

$$\frac{dS}{dt} = \mu_s X$$

where S is the substrate, μ_s the uptake rate and X the CDW at the given time. The uptake rates at 21 hours after inoculation seems to be too low for M145 and is caused by a too low estimate of the CDW. The uptake rates for glucose and given for M145 and M1152 in **Table S1** and **Table S2**, respectively.

Undecylprodigiosin (RED) production rate

We used the same method as for the uptake rates of glucose and glutamic acid to estimate the production rate of undecylprodigiosin (RED). The M1152 did not produce any RED (as expected). However, for the M145 the amount of undecylprodigiosin was measured both using MS (mass spectrometry) and OD (optical density). From the MS data it looks like the production of RED stops after approximately 53 hours, but from the OD measurements we observe a continuous increase until the end of the experiment. The hypothesis is that RED is continuously degraded into derivatives which are still measurable using OD but not using MS because of the different masses of the derivatives. Therefore, we have extrapolated the production rate of RED using the timepoints between 40 and 53 hours to estimate the production rate of RED at 57 hours (**Figure S10I**).

Germicidin A and B production rates

Production rates of germicidin A and B were fitted using linear regression of the last 6 data points across all three biological replicates for M145 (**Table S1**). For M1152, only the measured concentrations at 45 and 65 hours after inoculation were used to estimate the production rate of germicidin A and B (**Table S2**).

memote snapshot report of Sco-GEM

The following report (next page) was prepared by running memote version 0.9.12 in the command-line from the directory of the cloned Sco-GEM repository with the command: *memote report snapshot --custom-tests ComplementaryScripts/tests*



Independent Section

Contains tests that are independent of the class of modeled organism, a model's complexity or types of identifiers that are used to describe its components. Parameterization or initialization of the network is not required. See readme for more details.

Consistency

Stoichiometric Consistency	37.9%
Mass Balance	84.6%
Charge Balance	96.8%
Metabolite Connectivity	100.0%
Unbounded Flux In Default Medium	84.9%
Sub Total	69%

Annotation - Metabolites

Presence of Metabolite Annotation

Metabolite Annotations Per Database

pubchem.compound	18.5%
kegg.compound	82.4%
seed.compound	1.6%
inchkey	0.0%
inchi	18.6%
chebi	80.3%
hmdb	0.0%
reactome	0.0%
metanetx.chemical	88.2%
bigg.metabolite	95.9%
biocyc	70.2%

Specific Section

Covers general statistics and specific aspects of a metabolic network that are not universally applicable. See readme for more details.

SBML

SBML Level and Version	SBML Level 3 Version 1
FBC enabled	true

Basic Information

Model Identifier	2,073
Total Metabolites	2,612
Total Reactions	1,778
Total Genes	2
Total Compartments	1,47
Metabolic Coverage	

Metabolite Information

Unique Metabolites	1,836
Duplicate Metabolites in Identical Compartments	0
Metabolites without Charge	0
Metabolites without Formula	0
Medium Components	22

Reaction Information

Purely Metabolic Reactions	2,011
Purely Metabolic Reactions with Constraints	1
Transport Reactions	325



kegg.compound	100.0%
seed.compound	100.0%
inchikey	0.0%
inchi	100.0%
chebi	99.9%
hmdb	0.0%
reactome	0.0%
metanex.chemical	99.8%
bigg.metabolite	100.0%
biocyc	99.4%
Uniform Metabolite Identifier Namespace	100.0%
Sub Total	79%

Annotation - Reactions

Presence of Reaction Annotation	100.0%
Reaction Annotations Per Database	Info
rhea	37.4%
kegg.reaction	52.4%
seed.reaction	0.0%
metanex.reaction	79.2%
bigg.reaction	93.6%
reactome	0.0%
ec-code	64.4%
brenda	0.0%
biocyc	44.3%
Reaction Annotation Conformity Per Database	Info

Reactions With Partially Identical Annotations	0.09
Duplicate Reactions	0.00
Reactions With Identical Genes	0.48

Gene-Protein-Reaction (GPR) Associations

Reactions without GPR	342
Fraction of Transport Reactions without GPR	0.26
Enzyme Complexes	199

Biomass

Biomass Reactions Identified	10
Biomass Consistency	Info
MISC_PSEUDO	Errored
CARBOHYDRATE_PSEUDO	Errored
PROTEIN_PSEUDO_IRNA	Errored
LIPID_PSEUDO	Errored
CELL_WALL_PSEUDO	Errored
BIOMASS_SCO_IRNA	Errored
DNA_PSEUDO	Errored
RNA_PSEUDO	Errored
BIOMASS_SCO	Errored
PROTEIN_PSEUDO	Errored
Biomass Production In Default Medium	Info
MISC_PSEUDO	0.07
CARBOHYDRATE_PSEUDO	0.07
PROTEIN_PSEUDO_IRNA	0.07
LIPID_PSEUDO	0.07



seed_reaction	0.0%	>
metanetx_reaction	100.0%	>
bigg_reaction	100.0%	>
reactome	100.0%	>
ec_code	97.2%	>
brenda	0.0%	>
biocyc	100.0%	>
Uniform Reaction Identifier Namespace	100.0%	>

Sub Total 80% >

Annotation - Genes

Presence of Gene Annotation 89.5% >

Gene Annotations Per Database Info >

refseq	89.5%	>
uniprot	89.5%	>
ecogene	0.0%	>
kegg_genes	0.0%	>
ncbigi	0.0%	>
ncbigene	0.0%	>
ncbiprotein	0.0%	>
ccds	0.0%	>
hprid	0.0%	>
asap	0.0%	>
Gene Annotation Conformity Per Database	Info	>
refseq	11.1%	>
uniprot	100.0%	>

DNA_PSEUDO	0.07	>
RNA_PSEUDO	0.07	>
BIOMASS_SCO	0.07	>
PROTEIN_PSEUDO	0.07	>

Unrealistic Growth Rate In Default Medium Info >

MISC_PSEUDO	false	>
CARBOHYDRATE_PSEUDO	false	>
PROTEIN_PSEUDO_IRNA	false	>
LIPID_PSEUDO	false	>
CELL_WALL_PSEUDO	false	>
BIOMASS_SCO_IRNA	false	>
DNA_PSEUDO	false	>
RNA_PSEUDO	false	>
BIOMASS_SCO	false	>
PROTEIN_PSEUDO	false	>

Biomass Production In Complete Medium Info >

MISC_PSEUDO	123.45	>
CARBOHYDRATE_PSEUDO	123.45	>
PROTEIN_PSEUDO_IRNA	123.45	>
LIPID_PSEUDO	123.45	>
CELL_WALL_PSEUDO	123.45	>
BIOMASS_SCO_IRNA	123.45	>
DNA_PSEUDO	123.45	>
RNA_PSEUDO	123.45	>
BIOMASS_SCO	123.45	>
PROTEIN_PSEUDO	123.45	>



ncbigi	0.0%
ncbigene	0.0%
ncbiprotein	0.0%
ccds	0.0%
hprd	0.0%
asap	0.0%

Sub Total 40%

Annotation - SBO Terms

Metabolite General SBO Presence	100.0%
Metabolite SBO:0000247 Presence	99.6%
Reaction General SBO Presence	100.0%
Metabolic Reaction SBO:0000176 Presence	99.8%
Transport Reaction SBO:0000185 Presence	100.0%
Exchange Reaction SBO:0000627 Presence	100.0%
Demand Reaction SBO:0000628 Presence	100.0%
Sink Reactions SBO:0000632 Presence	Skipped
Gene General SBO Presence	89.5%
Gene SBO:0000243 Presence	89.5%
Biomass Reactions SBO:0000629 Presence	100.0%

Sub Total 89%

Total Score 77%

Total Score

CARBOHYDRATE_PSEUDO	0
PROTEIN_PSEUDO_IRNA	20
LIPID_PSEUDO	0
CELL_WALL_PSEUDO	0
BIOMASS_SCO_IRNA	0
DNA_PSEUDO	0
RNA_PSEUDO	0
BIOMASS_SCO	0
PROTEIN_PSEUDO	0

Blocked Biomass Precursors In Complete Medium

MISC_PSEUDO	0
CARBOHYDRATE_PSEUDO	0
PROTEIN_PSEUDO_IRNA	20
LIPID_PSEUDO	0
CELL_WALL_PSEUDO	0
BIOMASS_SCO_IRNA	0
DNA_PSEUDO	0
RNA_PSEUDO	0
BIOMASS_SCO	0
PROTEIN_PSEUDO	0

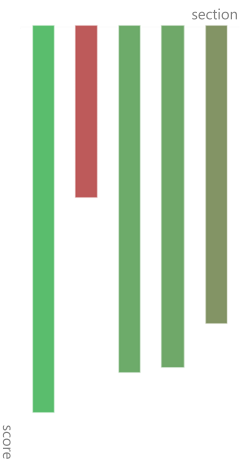
Ratio of Direct Metabolites in Biomass Reaction

MISC_PSEUDO	0.10
CARBOHYDRATE_PSEUDO	0.00
PROTEIN_PSEUDO_IRNA	0.00
LIPID_PSEUDO	0.00
CELL_WALL_PSEUDO	0.00



100%

Score per Category



RNA_PSEUDO

0.00

BIOMASS_SCO

0.00

PROTEIN_PSEUDO

0.00

Number of Missing Essential Biomass Precursors

Info

MISC_PSEUDO

29

CARBOHYDRATE_PSEUDO

36

PROTEIN_PSEUDO_IRNA

37

LIPID_PSEUDO

37

CELL_WALL_PSEUDO

1

BIOMASS_SCO_IRNA

1

DNA_PSEUDO

1

RNA_PSEUDO

1

BIOMASS_SCO

1

PROTEIN_PSEUDO

17

Energy Metabolism

Non-Growth Associated Maintenance Reaction

Growth-associated Maintenance in Biomass Reaction

MISC_PSEUDO

false

CARBOHYDRATE_PSEUDO

false

PROTEIN_PSEUDO_IRNA

false

LIPID_PSEUDO

false

CELL_WALL_PSEUDO

false

BIOMASS_SCO_IRNA

false

DNA_PSEUDO

false

RNA_PSEUDO

false

BIOMASS_SCO

false

Errored



Expand All

Readme

2019-10-04 20:34



Erroneous Energy-generating Cycles

	Info	
MNXM3	Skipped	>
MNXM63	Skipped	>
MNXM51	Skipped	>
MNXM121	Skipped	>
MNXM423	Skipped	>
MNXM6	Skipped	>
MNXM10	Skipped	>
MNXM38	Skipped	>
MNXM208	Skipped	>
MNXM191	Skipped	>
MNXM223	Skipped	>
MNXM7517	Skipped	>
MNXM12233	Skipped	>
MNXM558	Skipped	>
MNXM21	Skipped	>
MNXM89557	Skipped	>

Network Topology

Universally Blocked Reactions	683	>
Orphan Metabolites	140	>
Dead-end Metabolites	232	>
Stoichiometrically Balanced Cycles	156	>
Metabolite Production In Complete Medium	714	>
Metabolite Consumption In Complete Medium	876	>

Matrix Conditioning



Rank

1915



Degrees Of Freedom

697



Experimental Data Comparison

Growth Prediction

Skipped



Gene Essentiality Prediction

Skipped



Misc. Tests

Test if all metabolites have been given a name

1.00



Test CDA production

0.07



Test germicidinB production

0.37



Test growth for knockout-mutants from the transposon mutagenesis study by Xu et al.(2017)

0.76



Test germicidinA production

0.37



Test RED production

0.13



Test germicidinC production

0.33



Test that the growth rate is around 0.075

0.07



Test growth for knockout-mutants from the literature in given environments

0.63



Test growth for WT in given environments

0.96



Test if all reactions have been given a name

1.00



Test ACT production

Errored



Environment

Python Version
Platform
Memote Version

3.7.3
Windows
0.9.12

Supplemental references

Amara, A., Takano, E., and Breitling, R. (2018). Development and validation of an updated computational model of *Streptomyces coelicolor* primary and secondary metabolism. *BMC Genomics* *19*, 519.

Andrews, S. (2016). FastQC: a quality control tool for high throughput sequence data.

Bar-Even, A., Flamholz, A., Noor, E., and Milo, R. (2012). Thermodynamic constraints shape the structure of carbon fixation pathways. *Biochimica et Biophysica Acta (BBA) - Bioenergetics* *1817*, 1646–1659.

Battke, F., and Nieselt, K. (2011). Mayday SeaSight: combined analysis of deep sequencing and microarray data. *PLoS ONE* *6*, e16345.

Bordel, S., Agren, R., and Nielsen, J. (2010). Sampling the Solution Space in Genome-Scale Metabolic Networks Reveals Transcriptional Regulation in Key Enzymes. *PLOS Computational Biology* *6*, e1000859.

Bystrykh, L.V., Fernández-Moreno, M.A., Herrema, J.K., Malpartida, F., Hopwood, D.A., and Dijkhuizen, L. (1996). Production of actinorhodin-related “blue pigments” by *Streptomyces coelicolor* A3(2). *J. Bacteriol.* *178*, 2238–2244.

Caspi, R., Altman, T., Billington, R., Dreher, K., Foerster, H., Fulcher, C.A., Holland, T.A., Keseler, I.M., Kothari, A., Kubo, A., et al. (2014). The MetaCyc database of metabolic pathways and enzymes and the BioCyc collection of Pathway/Genome Databases. *Nucleic Acids Res* *42*, D459–D471.

Chicco, D., and Jurman, G. (2020). The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation. *BMC Genomics* *21*, 6.

Claessen, D., Rink, R., Jong, W. de, Siebring, J., Vreugd, P. de, Boersma, F.G.H., Dijkhuizen, L., and Wösten, H.A.B. (2003). A novel class of secreted hydrophobic proteins is involved in aerial hyphae formation in *Streptomyces coelicolor* by forming amyloid-like fibrils. *Genes Dev.* *17*, 1714–1726.

- Cokelaer, T., Pultz, D., Harder, L.M., Serra-Musach, J., and Saez-Rodriguez, J. (2013). BioServices: a common Python package to access biological Web Services programmatically. *Bioinformatics* 29, 3241–3242.
- Courtot, M., Juty, N., Knüpfer, C., Waltemath, D., Zhukova, A., Dräger, A., Dumontier, M., Finney, A., Golebiewski, M., Hastings, J., et al. (2011). Controlled vocabularies and semantics in systems biology. *Mol. Syst. Biol.* 7, 543.
- Distler, U., Kuharev, J., Navarro, P., Levin, Y., Schild, H., and Tenzer, S. (2014). Drift time-specific collision energies enable deep-coverage data-independent acquisition proteomics. *Nat. Methods* 11, 167–170.
- Ebrahim, A., Lerman, J.A., Palsson, B.O., and Hyduke, D.R. (2013). COBRApy: CONstraints-Based Reconstruction and Analysis for Python. *BMC Systems Biology* 7, 74.
- Elbourne, L.D.H., Tetu, S.G., Hassan, K.A., and Paulsen, I.T. (2017). TransportDB 2.0: a database for exploring membrane transporters in sequenced genomes from all domains of life. *Nucleic Acids Research* 45, D320–D324.
- Feist, A.M., Henry, C.S., Reed, J.L., Krummenacker, M., Joyce, A.R., Karp, P.D., Broadbelt, L.J., Hatzimanikatis, V., and Palsson, B.Ø. (2007). A genome-scale metabolic reconstruction for *Escherichia coli* K-12 MG1655 that accounts for 1260 ORFs and thermodynamic information. *Molecular Systems Biology* 3, 121.
- Flamholz, A., Noor, E., Bar-Even, A., and Milo, R. (2012). eQuilibrator—the biochemical thermodynamics calculator. *Nucleic Acids Res* 40, D770–D775.
- Fritzemeier, C.J., Hartleb, D., Szappanos, B., Papp, B., and Lercher, M.J. (2017). Erroneous energy-generating cycles in published genome scale metabolic networks: Identification and removal. *PLOS Computational Biology* 13, e1005494.
- Gomez-Escribano, J.P., and Bibb, M.J. (2011). Engineering *Streptomyces coelicolor* for heterologous expression of secondary metabolite gene clusters. *Microbial Biotechnology* 4, 207–215.

- Gubbens, J., Janus, M., Florea, B.I., Overkleeft, H.S., and van Wezel, G.P. (2012). Identification of glucose kinase-dependent and -independent pathways for carbon control of primary metabolism, development and antibiotic production in *Streptomyces coelicolor* by quantitative proteomics. *Molecular Microbiology* 86, 1490–1507.
- Haraldsdóttir, H.S., Cousins, B., Thiele, I., Fleming, R.M.T., and Vempala, S. (2017). CHRR: coordinate hit-and-run with rounding for uniform sampling of constraint-based models. *Bioinformatics* 33, 1741–1743.
- Hu, H., Zhang, Q., and Ochi, K. (2002). Activation of Antibiotic Biosynthesis by Specified Mutations in the *rpoB* Gene (Encoding the RNA Polymerase β Subunit) of *Streptomyces lividans*. *Journal of Bacteriology* 184, 3984–3991.
- Huang, D.W., Sherman, B.T., and Lempicki, R.A. (2009a). Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists. *Nucleic Acids Res.* 37, 1–13.
- Huang, D.W., Sherman, B.T., and Lempicki, R.A. (2009b). Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nat Protoc* 4, 44–57.
- Jeske, L., Placzek, S., Schomburg, I., Chang, A., and Schomburg, D. (2019). BRENDA in 2019: a European ELIXIR core data resource. *Nucleic Acids Res* 47, D542–D549.
- Kanehisa, M. (2000). KEGG: Kyoto Encyclopedia of Genes and Genomes. *Nucleic Acids Research* 28, 27–30.
- Kanehisa, M., Sato, Y., Furumichi, M., Morishima, K., and Tanabe, M. (2019). New approach for understanding genome variations in KEGG. *Nucleic Acids Research* 47, D590–D595.
- Karp, P.D., Latendresse, M., Paley, S.M., Kruppenacker, M., Ong, Q.D., Billington, R., Kothari, A., Weaver, D., Lee, T., Subhraveti, P., et al. (2016). Pathway Tools version 19.0 update: software for pathway/genome informatics and systems biology. *Brief Bioinform* 17, 877–890.

Karp, P.D., Billington, R., Caspi, R., Fulcher, C.A., Latendresse, M., Kothari, A., Keseler, I.M., Krummenacker, M., Midford, P.E., Ong, Q., et al. (2019). The BioCyc collection of microbial genomes and metabolic pathways. *Briefings in Bioinformatics* 20, 1085–1093.

Kaufman, D.E., and Smith, R.L. (1998). Direction Choice for Accelerated Convergence in Hit-and-Run Sampling. *Operations Research* 46, 84–95.

Keesey, J.K., Bigelis, R., and Fink, G.R. (1979). The product of the *his4* gene cluster in *Saccharomyces cerevisiae*. A trifunctional polypeptide. *J. Biol. Chem.* 254, 7427–7433.

Kieser, T., Bibb, M.J., Buttner, M.J., Chater, K.F., and Hopwood, D.A. (2000). *Practical Streptomyces Genetics* (Norwich, UK: John Innes Foundation).

Kim, D., Langmead, B., and Salzberg, S.L. (2015). HISAT: a fast spliced aligner with low memory requirements. *Nat. Methods* 12, 357–360.

King, Z.A., Lu, J., Dräger, A., Miller, P., Federowicz, S., Lerman, J.A., Ebrahim, A., Palsson, B.O., Lewis, N.E., and J., H. (2016). BiGG Models: A platform for integrating, standardizing and sharing genome-scale models. *Nucleic Acids Research* 44, D515–D522.

Kumelj, T., Sulheim, S., Wentzel, A., and Almaas, E. (2019). Predicting Strain Engineering Strategies Using iKS1317: A Genome-Scale Metabolic Model of *Streptomyces coelicolor*. *Biotechnol. J.* 14, 1800180.

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., Durbin, R., and 1000 Genome Project Data Processing Subgroup (2009). The Sequence Alignment/Map format and SAMtools. *Bioinformatics* 25, 2078–2079.

Liao, Y., Smyth, G.K., and Shi, W. (2014). featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. *Bioinformatics* 30, 923–930.

Lieven, C., Beber, M.E., Olivier, B.G., Bergmann, F.T., Ataman, M., Babaei, P., Bartell, J.A., Blank, L.M., Chauhan, S., Correia, K., et al. (2018). Memote: A community-driven effort towards a standardized genome-scale metabolic model test suite. *BioRxiv* 350991.

- Love, M.I., Huber, W., and Anders, S. (2014). Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol.* *15*, 550.
- Maurer, K.H., Pfeiffer, F., Zehender, H., and Mecke, D. (1983). Characterization of two glyceraldehyde-3-phosphate dehydrogenase isoenzymes from the pentalenolactone producer *Streptomyces arenae*. *J. Bacteriol.* *153*, 930–936.
- Megchelenbrink, W., Huynen, M., and Marchiori, E. (2014). optGpSampler: An Improved Tool for Uniformly Sampling the Solution-Space of Genome-Scale Metabolic Networks. *PLOS ONE* *9*, e86587.
- Michael Waskom, Olga Botvinnik, Drew O’Kane, Paul Hobson, Joel Ostblom, Saulius Lukauskas, David C Gemperline, Tom Augspurger, Yaroslav Halchenko, John B. Cole, et al. (2018). mwaskom/seaborn: v0.9.0 (July 2018) (Zenodo).
- Moretti, S., Martin, O., Van Du Tran, T., Bridge, A., Morgat, A., and Pagni, M. (2016). MetaNetX/MNXref – reconciliation of metabolites and biochemical reactions to bring together genome-scale metabolic networks. *Nucleic Acids Res* *44*, D523–D526.
- NCBI Resource Coordinators (2017). Database Resources of the National Center for Biotechnology Information. *Nucleic Acids Research* *45*, D12–D17.
- Nieselt, K., Battke, F., Herbig, A., Bruheim, P., Wentzel, A., Jakobsen, Ø.M., Sletta, H., Alam, M.T., Merlo, M.E., Moore, J., et al. (2010). The dynamic architecture of the metabolic switch in *Streptomyces coelicolor*. *BMC Genomics* *11*, 10.
- Noor, E. (2018). Removing both Internal and Unrealistic Energy-Generating Cycles in Flux Balance Analysis. ArXiv:1803.04999 [q-Bio].
- Noor, E., Haraldsdóttir, H.S., Milo, R., and Fleming, R.M.T. (2013). Consistent Estimation of Gibbs Energy Using Component Contributions. *PLOS Computational Biology* *9*, e1003098.
- Okonechnikov, K., Conesa, A., and García-Alcalde, F. (2016). Qualimap 2: advanced multi-sample quality control for high-throughput sequencing data. *Bioinformatics* *32*, 292–294.

Rappsilber, J., Mann, M., and Ishihama, Y. (2007). Protocol for micro-purification, enrichment, pre-fractionation and storage of peptides for proteomics using StageTips. *Nat Protoc* 2, 1896–1906.

Rauch, G., Ehammer, H., Bornemann, S., and Macheroux, P. (2008). Replacement of two invariant serine residues in chorismate synthase provides evidence that a proton relay system is essential for intermediate formation and catalytic activity: Proton relay system in chorismate synthase. *FEBS Journal* 275, 1464–1473.

Saier, M.H., Reddy, V.S., Tsu, B.V., Ahmed, M.S., Li, C., and Moreno-Hagelsieb, G. (2016). The Transporter Classification Database (TCDB): recent advances. *Nucleic Acids Res* 44, D372–D379.

Sánchez, B.J., Zhang, C., Nilsson, A., Lahtvee, P.-J., Kerkhoven, E.J., and Nielsen, J. (2017). Improving the phenotype predictions of a yeast genome-scale metabolic model by incorporating enzymatic constraints. *Molecular Systems Biology* 13, 935.

Schendel, F.J., Mueller, E., Stubbe, J., Shiau, A., and Smith, J.M. (1989). Formylglycinamide ribonucleotide synthetase from *Escherichia coli*: cloning, sequencing, overproduction, isolation, and characterization. *Biochemistry* 28, 2459–2471.

Szklarczyk, D., Gable, A.L., Lyon, D., Junge, A., Wyder, S., Huerta-Cepas, J., Simonovic, M., Doncheva, N.T., Morris, J.H., Bork, P., et al. (2019). STRING v11: protein-protein association networks with increased coverage, supporting functional discovery in genome-wide experimental datasets. *Nucleic Acids Res.* 47, D607–D613.

Talfournier, F., Stines-Chaumeil, C., and Branlant, G. (2011). Methylmalonate semialdehyde dehydrogenase from *Bacillus subtilis*: substrate specificity and coenzyme A binding. *J. Biol. Chem.* jbc.M110.213280.

The UniProt Consortium (2019). UniProt: a worldwide hub of protein knowledge. *Nucleic Acids Res* 47, D506–D515.

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

Virtanen, P., Gommers, R., Oliphant, T.E., Haberland, M., Reddy, T., Cournapeau, D., Burovski, E., Peterson, P., Weckesser, W., Bright, J., et al. (2020). SciPy 1.0: fundamental algorithms for scientific computing in Python. *Nat Methods* *17*, 261–272.

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

Wentzel, A., Bruheim, P., Øverby, A., Jakobsen, Ø.M., Sletta, H., Omara, W.A.M., Hodgson, D.A., and Ellingsen, T.E. (2012). Optimized submerged batch fermentation strategy for systems scale studies of metabolic switching in *Streptomyces coelicolor* A3(2). *BMC Systems Biology* *6*, 59.

Wessel, D., and Flügge, U.I. (1984). A method for the quantitative recovery of protein in dilute solution in the presence of detergents and lipids. *Anal. Biochem.* *138*, 141–143.

Paper 5

Dynamic allocation of carbon storage and nutrient-dependent exudation in a revised genome-scale model of *Prochlorococcus*

Shany Ofaim, Snorre Sulheim, Eivind Almaas, Daniel Sher and Daniel Segrè.

Frontiers in Genetics, 12, 91 (2021).



Dynamic Allocation of Carbon Storage and Nutrient-Dependent Exudation in a Revised Genome-Scale Model of *Prochlorococcus*

Shany Ofaim^{1,2†}, Snorre Sulheim^{1,3,4†}, Eivind Almaas^{3,5}, Daniel Sher² and Daniel Segrè^{1,6,7,8*}

OPEN ACCESS

Edited by:

Karoline Faust,
KU Leuven, Belgium

Reviewed by:

Abhinav Achreja,
University of Michigan, United States
Adam Martiny,
University of California, Irvine,
United States
Steffen Waldherr,
KU Leuven, Belgium

*Correspondence:

Daniel Segrè
dsegrè@bu.edu

[†] These authors have contributed
equally to this work

Specialty section:

This article was submitted to
Systems Biology,
a section of the journal
Frontiers in Genetics

Received: 22 July 2020

Accepted: 14 January 2021

Published: 09 February 2021

Citation:

Ofaim S, Sulheim S, Almaas E,
Sher D and Segrè D (2021) Dynamic
Allocation of Carbon Storage
and Nutrient-Dependent Exudation
in a Revised Genome-Scale Model
of *Prochlorococcus*.
Front. Genet. 12:586293.
doi: 10.3389/fgene.2021.586293

¹ Bioinformatics Program and Biological Design Center, Boston University, Boston, MA, United States, ² Department of Marine Biology, University of Haifa, Haifa, Israel, ³ Department of Biotechnology and Food Science, NTNU – Norwegian University of Science and Technology, Trondheim, Norway, ⁴ Department of Biotechnology and Nanomedicine, SINTEF Industry, Trondheim, Norway, ⁵ K.G. Jebsen Center for Genetic Epidemiology, NTNU – Norwegian University of Science and Technology, Trondheim, Norway, ⁶ Department of Biomedical Engineering, Boston University, Boston, MA, United States, ⁷ Department of Physics, Boston University, Boston, MA, United States, ⁸ Department of Biology, Boston University, Boston, MA, United States

Microbial life in the oceans impacts the entire marine ecosystem, global biogeochemistry and climate. The marine cyanobacterium *Prochlorococcus*, an abundant component of this ecosystem, releases a significant fraction of the carbon fixed through photosynthesis, but the amount, timing and molecular composition of released carbon are still poorly understood. These depend on several factors, including nutrient availability, light intensity and glycogen storage. Here we combine multiple computational approaches to provide insight into carbon storage and exudation in *Prochlorococcus*. First, with the aid of a new algorithm for recursive filling of metabolic gaps (ReFill), and through substantial manual curation, we extended an existing genome-scale metabolic model of *Prochlorococcus* MED4. In this revised model (*i*SO595), we decoupled glycogen biosynthesis/degradation from growth, thus enabling dynamic allocation of carbon storage. In contrast to standard implementations of flux balance modeling, we made use of forced influx of carbon and light into the cell, to recapitulate overflow metabolism due to the decoupling of photosynthesis and carbon fixation from growth during nutrient limitation. By using random sampling in the ensuing flux space, we found that storage of glycogen or exudation of organic acids are favored when the growth is nitrogen limited, while exudation of amino acids becomes more likely when phosphate is the limiting resource. We next used COMETS to simulate day-night cycles and found that the model displays dynamic glycogen allocation and exudation of organic acids. The switch from photosynthesis and glycogen storage to glycogen depletion is associated with a redistribution of fluxes from the Entner–Doudoroff to

the Pentose Phosphate pathway. Finally, we show that specific gene knockouts in *iSO595* exhibit dynamic anomalies compatible with experimental observations, further demonstrating the value of this model as a tool to probe the metabolic dynamic of *Prochlorococcus*.

Keywords: constraint-based reconstruction and analysis (COBRA), flux balance analysis (FBA), computation of microbial ecosystems in time and space (COMETS), cyanobacteria, exudation, gap-filling algorithm, photosynthesis

INTRODUCTION

Marine phytoplankton perform about one-half of the photosynthesis on Earth (Field et al., 1998). *Prochlorococcus* is one of the most abundant phytoplankton clades in the world's oceans and is estimated to produce about 4 Gt of organic carbon annually (Flombaum et al., 2013). As such, these clades play a key role in a variety of ecosystems (Partensky and Garczarek, 2010; Biller et al., 2015). Recent evolutionary studies suggested several evolved metabolic innovations contributing to high picocyanobacterial abundance in the harsh oligotrophic ocean waters, usually limited by several nutrients such as nitrogen, phosphorus, and iron. These innovations include a proteome that contains less nitrogen rich amino acids (Gilbert and Fagan, 2011), membranes that contain glyco- and sulfolipids rather than phospholipids (Van Mooy et al., 2006) and streamlining of the genome associated with outsourcing of important cellular functions to co-occurring organisms (Holtendorff et al., 2008; Partensky and Garczarek, 2010; Morris et al., 2012; Ma L. et al., 2018; Braakman, 2019).

Another innovation employed by these organisms is an increased metabolic rate that in turn manifest in the exudation of organic compounds (Fogg et al., 1965; Mague et al., 1980; López-Sandoval et al., 2013; Braakman et al., 2017; Braakman, 2019; Moran and Durham, 2019). Typically, 2–25% of the carbon fixed by photosynthesis is released by exudation from the cell, although values as high as 90% have been reported (Bertilsson et al., 2005; López-Sandoval et al., 2013; Rothrosenberg et al., 2019; Szul et al., 2019). This exudation, combined with cell death, lytic viral infections, and grazing debris made by predators (“sloppy feeding”), makes dissolved organic matter of phytoplankton origin omnipresent in natural waters (Thornton, 2014). However, it is currently impossible to provide a universal chemical description of dissolved organic matter (Kujawinski, 2011; Arrieta et al., 2015; Moran et al., 2016), partly because the exuded organic compounds differ between strains and environmental conditions (Becker et al., 2014; Ma X. et al., 2018). Nevertheless, in general, phytoplankton exudate includes a small proportion of low-molecular weight compounds, such as organic acids, carbohydrates, and amino acids (Bertilsson et al., 2005), as well as a larger proportion of complex, high-molecular weight compounds (Kujawinski, 2011). Another strategy employed by these bacteria to manage their carbon budget is the internal storage of carbon in polymeric form, specifically, glycogen (Zinser et al., 2009; Reimers et al., 2017; Luan et al., 2019). The extent to which *Prochlorococcus*, in particular, also stores glycogen has recently been measured,

showing increased glycogen pools (up to 40 fg cell⁻¹) in nitrogen-limited conditions compared to nitrogen-replete (Szul et al., 2019). Glycogen accumulates in the bacterial cell during the light hours and was recently suggested to have two primary roles; as energy storage in preparation for darkness and as a regulation strategy to manage high-light photosynthesis products (Welkie et al., 2019). The allocation of glycogen is suggested to be tightly associated with the overflow metabolism hypothesis and also known to be widely affected by nutrient limitations (Damrow et al., 2016; Cano et al., 2018; Forchhammer and Schwarz, 2019; Szul et al., 2019). Importantly, the carbon fixed and released by phytoplankton is then used by heterotrophic organisms as a source of energy, whereas the heterotrophic bacteria may recycle nutrient elements and support the growth of phytoplankton in other ways, as suggested by the Black Queen Hypothesis (Amin et al., 2012; Morris et al., 2012; Moran et al., 2016; Cirri and Pohnert, 2019; Moran and Durham, 2019). Thus, carbon fixation, storage and release are tightly intertwined with microbial interactions and microbial ecosystem dynamics.

Quantitative models at various scales have provided critical insights into how ocean microbial ecosystems function, and how they are related to broader biogeochemical cycles (Deutsch et al., 2007; Follows et al., 2007; Arteaga et al., 2016; Coles et al., 2017; Foster et al., 2018; Moradi et al., 2018; Nicholson et al., 2018; Braakman, 2019; Oschlies et al., 2019; Ward et al., 2019). Most of these models represent organisms in terms of simplified stoichiometric reactions converting elements into biomass, thus making it possible to incorporate biological processes into dynamic-coupled Earth System models (Follows et al., 2007; Reid, 2012). The exponential increase in genomic information on marine organisms provides an opportunity to seek methods to link such detailed genome-scale information to biochemical flows (Coles et al., 2017). In recent years, genome-scale metabolic models (GEMs), combined with linear programming, have made it possible to produce testable predictions of metabolic phenotypes of individual organisms or microbial communities (Gu et al., 2019). This computational framework is based on the identification of individual enzymes and transporters in an organism's genome, and on simplifying assumptions that bypass the need for kinetic parameters (Maarleveld et al., 2013; O'Brien et al., 2015; Casey et al., 2016; Kim et al., 2016; Reimers et al., 2017). While genome-scale modeling has proven to be a powerful approach in cyanobacterial model organisms such as *Synechocystis* sp. PCC 6803, *Synechococcus elongatus* PCC 7942 and *Prochlorococcus* MED4 (Kettler et al., 2007; Knoop et al., 2013; Broddrick et al., 2016; Casey et al., 2016; Yoshikawa et al., 2017), the exudation of organic compounds in

phototrophic organisms has not been studied in detail through Flux Balance Analysis (FBA) or similar methods (Varma and Palsson, 1994; Orth et al., 2010). On the other hand, several examples exist of FBA-based predictions of exudation-mediated interactions between different species, including those generated using the Computational of Microbial Ecosystems in Time and Space (COMETS) platform (Harcombe et al., 2014). In fact, FBA calculations also suggest that “costless” secretions (i.e., secretions that do not induce a fitness cost) might be quite common, and can support the growth of co-occurring organisms (Pacheco et al., 2018).

Experimental evidence and theoretical considerations indicate that *Prochlorococcus* exudes different metabolites in a way that strongly depends on environmental conditions (Dubinsky and Berman-Frank, 2001; Szul et al., 2019) as well as on the strain’s genetic makeup (Becker et al., 2014; Roth-rosenberg et al., 2019). While GEMs can be used to predict these fluxes, they require modifications to deal with processes not usually considered in FBA, including: (a) the special nature of photon fluxes [which, unlike molecular fluxes, cannot easily be “shut off” at short time scales (Dubinsky and Berman-Frank, 2001)]; (b) the buffering role of intracellular storage molecules such as glycogen. The primary focus of this study is to obtain better knowledge of the potential metabolic effect of a combination of key nutrients (carbon, nitrogen, phosphorous, and light) and carbon fixation rate on the allocation (including storage and exudation) of carbon in *Prochlorococcus* using a revised genome-scale metabolic model (Figure 1). We start by describing model revisions and updates to capture the current, most complete metabolic knowledge available for *Prochlorococcus*. Next, we use a variety of FBA approaches to uncover the potential relationships between a set of key nutrients, carbon storage and exudates in static and dynamic (time dependent) settings. The implementation and use of these approaches improve our understanding of the intricate metabolic workings of *Prochlorococcus* and provide insights on its storage and exudation trends under different environmental conditions.

MATERIALS AND METHODS

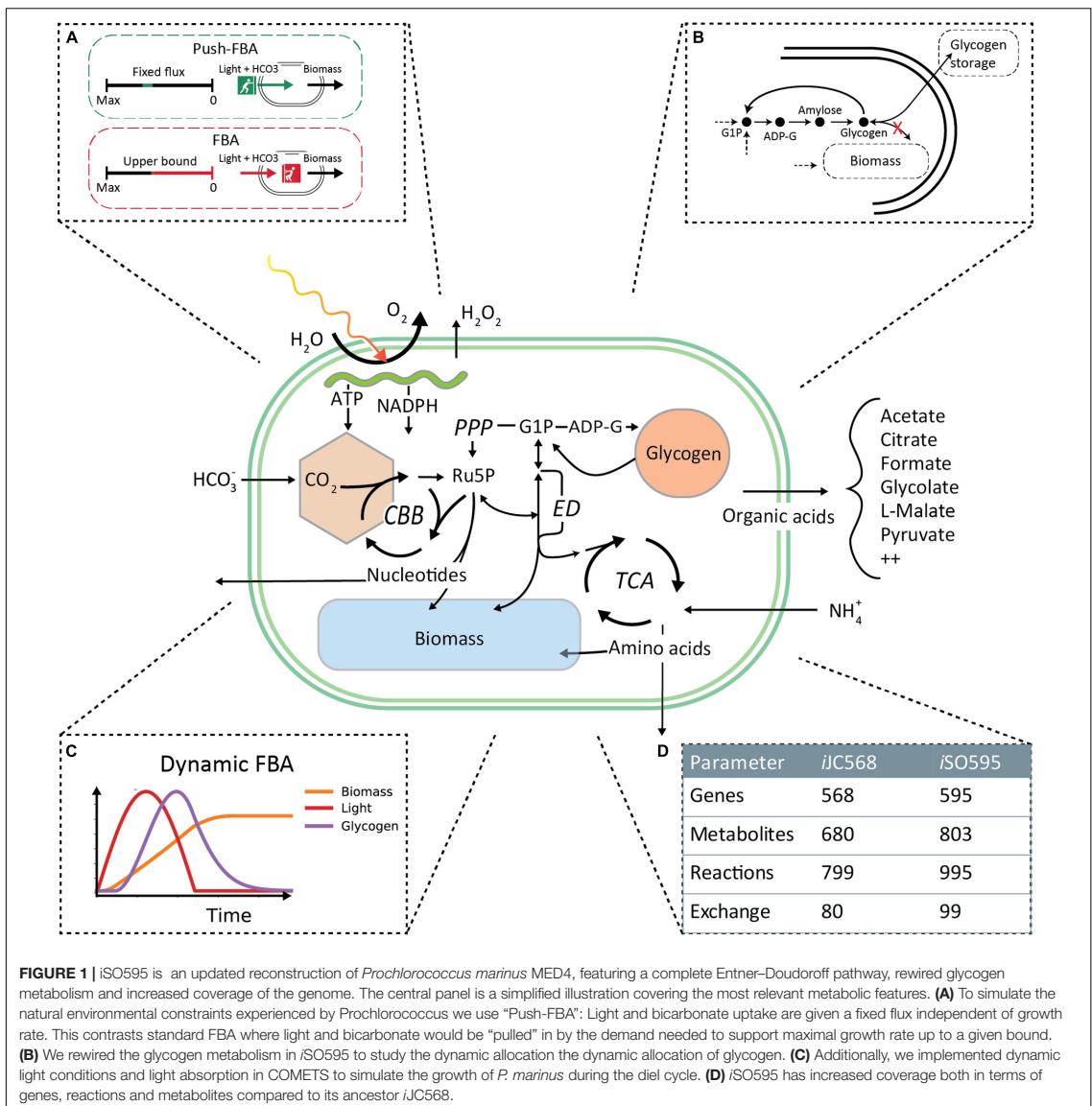
Model Update and Curation

The *iJC568* genome-scale reconstruction of *Prochlorococcus marinus subsp. pastoris str. CMP1986* (referred to throughout the manuscript as MED4) as described by Casey et al. (2016), was used as the starting point for model enhancement. The update process started with an in-depth study of the reconstructed network and available knowledge not previously incorporated into the model of the organism. During this process, we ended up implementing the following specific steps of curation and update: (i) A key modification to the model was the decoupling between the glycogen storage flux and the biomass production. In standard stoichiometric reconstructions for FBA modeling (Thiele et al., 2011; Nogales et al., 2012; Feist et al., 2014; Broddrick et al., 2016; Monk et al., 2017; Kavvas et al., 2018), glycogen is listed as one of the biomass components, thus accounting for the carbon flux into storage. However, given the fixed stoichiometry of biomass composition, this classical

implementation cannot account for the time-dependent storage and re-utilization of glycogen observed in picocyanobacteria. We thus removed the glycogen from the biomass function and streamlined the existing glycogen granule representation to a direct link between ADP-Glucose to the production of glycogen (Figure 1B). (ii) In addition to targeted refinement of selected reactions, we used the KEGG database (Kanehisa and Goto, 2000) to perform an extensive search for previously known but missing metabolic reaction annotations. Indeed, we found 354 reactions that could be potentially added to the existing network. To incorporate this knowledge, we developed a semi-automated algorithm (ReFill, described below). (iii) We coupled the implementation of the algorithm with several steps of manual curation. These included the addition of transports, such as that of hydrogen peroxide and ethanol, known to diffuse across the cell membrane (Seaver and Imlay, 2001; Noreña-Caro and Benton, 2018), and the addition of the complete Entner-Doudoroff pathway, that has recently been discovered in cyanobacteria (Chen et al., 2016). Additionally, we performed a BLAST search (Supplementary Material 1) (Altschul et al., 1990) from which we identified 6PG-dehydratase (EC: 4.2.1.12) encoded by PMM0774, thereby completing this pathway in the model reconstruction. (iv) The revised model was checked for redox and elemental balance. Since the biomass function was based on experimental data (Casey et al., 2016), it was not updated. In line with best practices, a *memote* quality assessment (Supplementary Material 2) (Lieven et al., 2020), as well as model files and a detailed changelog, are provided at https://github.com/segrelab/Prochlorococcus_Model. All reactions added to *iJC568* to form *iSO595* are found in Supplementary Table 1 and modified reactions are found in Supplementary Table 2.

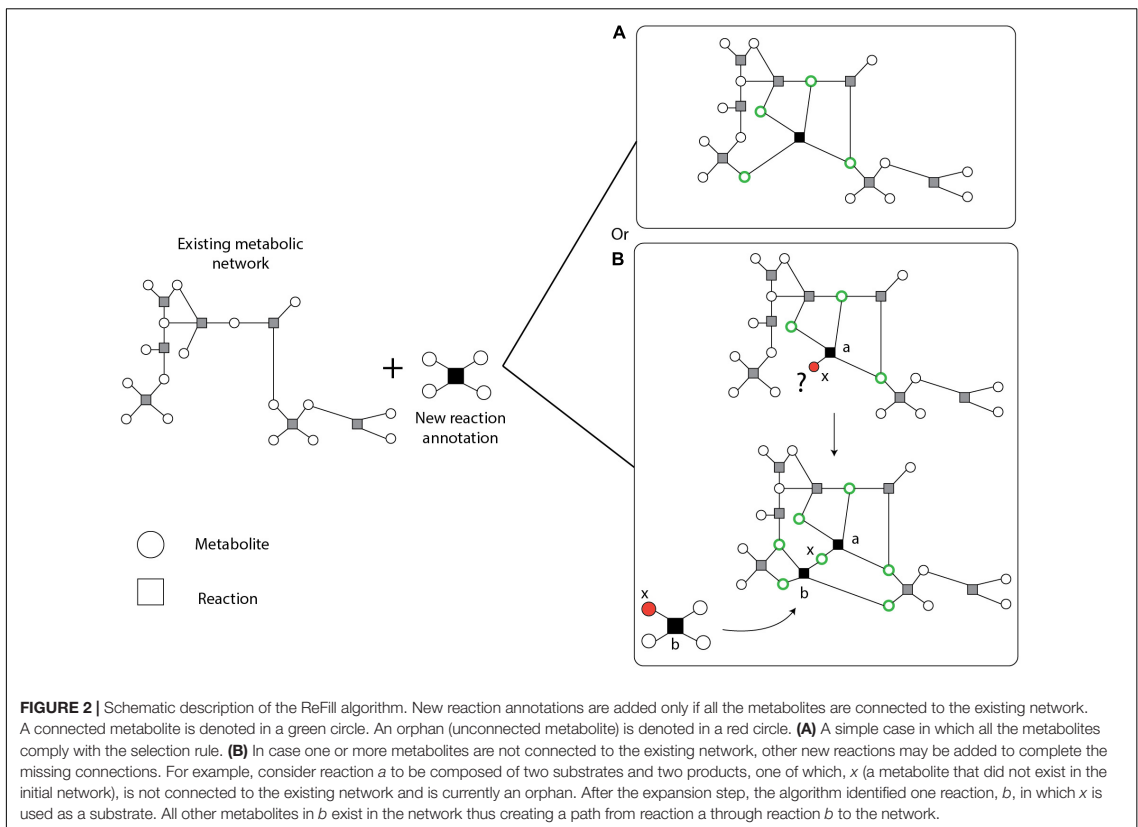
ReFill Algorithm

Following an extensive search of literature and the KEGG (Kanehisa and Goto, 2000), TransportDB (Elbourne et al., 2017) and Metabolights (Haug et al., 2013) databases, we found a large number of new or previously known but missing reaction, transporter, and metabolite annotations. Adding large amounts of data to an existing network might create new gaps and may give rise to new blocked reactions and orphan metabolites that in general reduce model quality and can convolute later curation efforts, quality control or assessment of model predictions. To add this knowledge to the network in a controlled approach, we developed the semi-automated recursive algorithm ReFill (Recursive Filler of metabolic gaps). The algorithm is based on the principle of using diverse information, such as enzyme and reaction annotations, and experimental data (such as metabolomics), to selectively increase the metabolic knowledge of an organism’s existing curated genome-scale metabolic network. ReFill makes use of a repository of reactions, in this case KEGG reaction annotations for MED4 absent from the model, to construct all potential chains of reactions connecting two metabolites in the existing network. It systematically tests the potential of adding each new reaction and suggests adding it only if it can be a part of a chain in which all the metabolites are part of a path in the network (Figure 2). This prevents



the creation of new orphan metabolites and potential blocked reactions. The algorithm starts by selecting a reaction from the repository. It then inspects each metabolite in the reaction for presence in the existing network. In case a metabolite is not present, the set of available reactions is scanned for other reactions using this metabolite as a substrate or product. If such a reaction is found, it is added to the chain of potential reactions. The algorithm then iteratively expands the chain until either the repository is exhausted or all the metabolites in the most recent reaction added are present in the network. After all the

possible chains of new reactions are expanded, the algorithm examines the connectivity of all the metabolites in each chain (see example in **Figure 2B**). Following the manual addition of transporters found through TransportDB (Elbourne et al., 2017) and Metablights (Haug et al., 2013) (Study MTBLS567), using the ReFill algorithm, we updated reactions that belong to several different pathways, including metabolism of cofactors and vitamins, carbohydrate metabolism, amino acid metabolism and nucleotide metabolism. A complete list of added reactions can be found in **Supplementary Table 1**. ReFill was coded in python



3.7 and generates MATLAB-compatible files formatted to be used with the COBRA Toolbox (Heirendt et al., 2019), including a list of suggested reactions to add and their gene-reaction rules. Other outputs include the added reaction chains and possible metabolic circuits that can be formed by these additions.

Parameter Sampling

To study the effects of combinations of key nutrients on glycogen production and exudation in the *i*SO595 model we focused on four parameters representing the uptake fluxes of light, bicarbonate, ammonium, and phosphate. Light and inorganic carbon (bicarbonate) are the substrates for photosynthesis, whereas nitrogen and phosphorus limit the growth of *Prochlorococcus* in large regions of the world ocean (Davey et al., 2008; Moore et al., 2013; Saito et al., 2014), and nutrient limitation is likely to influence the exudation of fixed carbon (Dubinsky and Berman-Frank, 2001). We sampled 10,000 different environmental conditions by drawing random values from uniform distributions of these four parameters. The range of each parameter was based on physiologically relevant ranges we extracted from the literature and on the requirement that each range covers important phase transitions, such as nutrient and light limitations (**Supplementary Table 3**).

Light flux was converted from micromole quanta $m^{-2} sec^{-1}$ to $mmol\ gDW^{-1}h^{-1}$ similarly to Nogales et al. (2012) using 8% photosynthesis efficiency rate (Zinser et al., 2009). All uptake flux parameters are described in FBA-compatible units ($mmol\ gDW^{-1}h^{-1}$), while corresponding values in biogeochemistry relevant units are illustrated in **Supplementary Table 3**. As *MED4* is a photoautotroph, it is exposed to a constant stream of light during daylight hours. The bacterium is then forced, or ‘pushed,’ to fix carbon even when there is not enough of other elements, such as nitrogen or phosphate, to combine the fixed carbon into biomass. To capture this phenomenon *in silico* we developed a ‘push’-FBA framework where we fixed both upper and lower uptake rates of light and bicarbonate (**Figure 1A**). For the other sampled nutrients, ammonium and phosphate, we defined standard FBA bounds where the maximal uptake rate was set to the sampled value and the lower bound was set to zero. Note that we only considered uptake of sulfur in the form of sulfate (not hydrogen sulfide), and no upper limit was set for the uptake of sulfate because of its abundance in seawater. The maximum rate of RuBisCO (R00024) was fixed to $4.7\ mmol\ gDW^{-1}\ h^{-1}$, as previously reported (Casey et al., 2016). Before sampling we blocked a set of artificial exchange reactions that were added in the previous version of the model, most likely

to allow export of dead-end metabolites that would otherwise limit flux feasibility (Supplementary Table 4). Subsequently, we removed all unconditionally blocked reactions in the model to speed up computations. For each random sample, we first tested the model for feasibility using FBA (Varma and Palsson, 1994). If the solver returned a solution that was feasible and optimal, we further calculated optimal fluxes with parsimonious FBA (Lewis et al., 2010), and determined the range of possible fluxes at optimum with Flux Variability Analysis (FVA) (Gudmundsson and Thiele, 2010). Exchange fluxes from FBA, parsimonious FBA and FVA were recorded and used in subsequent analyses. All environmental sampling and calculations were performed using CobraPy (Ebrahim et al., 2013) and GUROBI 8.1.1 (Gurobi Optimization, Inc., Houston, TX, United States).

Statistical Analysis of the Sampled Spaces

We sampled 10,000 different environmental conditions based on the flux ranges described above, and analyzed the results of FBA optimization, with the goal of characterizing the distribution of, and correlation between, specific exchange (import/export) fluxes. To that end, we calculated Pearson correlations between exchange reaction fluxes in the sampling data using the python (version 3.7) Pandas package version 1.0.3 (McKinney, 2010). While negative values are normally used to define uptake in FBA, we converted them to positive values for the uptake of light, bicarbonate, phosphate, ammonium, and sulfate when calculating correlations to ease interpretation of the results. We also performed hierarchical clustering using the Nearest Point Algorithm in SciPy (Virtanen et al., 2020) to sort the order of the compounds in the correlation matrix.

We performed dimensionality reduction on normalized exchange reaction fluxes using the T-distributed Stochastic Neighbor Embedding (t-SNE) method (van der Maaten and Hinton, 2008) in Scikit-learn (Pedregosa et al., 2011) with perplexity of 50 and 3,000 iterations. The reaction fluxes were normalized to $[-1,1]$ by dividing by the maximum absolute flux value of each reaction to ensure a consistent influence on the t-SNE results from the different exchange reactions. We considered other normalization schemes, in particular standardization, but found that it was preferable not to center the data to easily discriminate uptake and exudation without further modifications in subsequent data visualization. Finally, the t-SNE transformed data was clustered using HDBSCAN (McInnes et al., 2017) with a minimum cluster size of 200. Transport of inorganic ions, water, and protons were not considered when calculating correlations, dimensionality reduction or clustering. We also discarded transport reactions with no absolute flux value above 10^{-3} mmol gDW $^{-1}$ h $^{-1}$ in any of the environmental samples.

Dynamic Modeling of Light Absorption During the Diel Cycle in COMETS

Cyanobacteria follow a diel cycle. To capture this dynamic behavior, we extended the Computation Of Microbial Ecosystems

in Time and Space (COMETS) platform (Harcombe et al., 2014; Dukovski et al., 2020), and developed a module for diurnal-cycle simulations allowing oscillations of light intensity and light absorption. Attenuation of light through each grid cell was modeled using the Beer–Lambert law, as described previously (Yang, 2011; Gomez et al., 2014):

$$I(t, z) = I_0(t)e^{-(a_w + a_{dw}X(t))z} \quad (1)$$

Here, $I(t, z)$ is the light irradiance given in mmol photons m $^{-2}$ s $^{-1}$, t is the time, z is the depth (from the top of the grid cell), a_{dw} is the cell- and wavelength-specific absorption coefficient given in m 2 gDW $^{-1}$, a_w the absorption coefficient of pure water given in m $^{-1}$, $X(t)$ the biomass concentration in gDW m $^{-3}$, and $I_0(t)$ the time-dependent incident light irradiance at the top of the grid cell. In the current version, we simplified the process by assuming that the light irradiance is either monochromatic or a sum of the total light bandwidth, and the absorption coefficient should match the wavelength(s) of the light source. The total light attenuation (ΔI) through a grid cell of thickness Δz is then

$$\Delta I(t) = I(t, 0) - I(t, \Delta z) = I_0 \left(1 - e^{-(a_w + a_{dw}X(t))\Delta z} \right) \quad (2)$$

The light absorbed by the cells is a fraction of the total light attenuation, i.e.,

$$I_{\text{abs}}(t) = \frac{\Delta I(t) \cdot a_{dw}X(t)}{a_w + a_{dw}X(t)}. \quad (3)$$

The total number of photons absorbed per dry cell weight $\Phi(t)$ in mmol photons gDW $^{-1}$ s $^{-1}$ by the cells within a grid cell of thickness Δz , volume V , and surface area A is then

$$\Phi(t) = \frac{I_{\text{abs}}(t) \cdot A}{X(t) \cdot V} = \frac{I_0(t)}{\Delta z} \frac{a_{dw}}{a_w + a_{dw}X(t)} \left(1 - e^{-(a_w + a_{dw}X(t))\Delta z} \right). \quad (4)$$

For all COMETS simulations presented here we have used monochromatic light at 680 nm with a calculated biomass-specific absorption coefficient a_{dw} as previously described (Morel and Bricaud, 1981; Bricaud et al., 2004). Briefly, the biomass-specific absorption is the weighted sum of the absorption coefficients of the light-absorbing pigments divinyl-chlorophyll A and B, since none of the other pigments in *Prochlorococcus* absorb light at 680 nm. Additionally, to account for the discrete distribution of chlorophyll into separate cells, the absorption coefficient is scaled by the packaging factor. All coefficients used to calculate light attenuation and absorption are provided in Table 1.

The changing light conditions throughout a diel cycle was modeled as

$$I_0(t) = A \max(\sin(\omega t), 0), \quad (5)$$

where the angular frequency is $\omega = \frac{2\pi}{T}$.

Following the development of the diel cycle simulation capability in COMETS we set out to dynamically simulate the growth of MED4. Since the nutrient uptake follows Michaelis–Menten kinetics, we estimated the kinetic parameters V_{max} and K_m using a heuristic approach from experimental data

TABLE 1 | Coefficients and values used to calculate light absorption in COMETS.

Symbol	Description	Value	Unit	Reference
λ	Wavelength	680	nm	
a_w	Absorption coefficient of water at 680 nm	0.465	m^{-1}	Pope and Fry, 1997
a_{dw}	Biomass-specific absorption coefficient	0.285	$m^2 \text{ gDW}^{-1}$	
$a_{dvchl-A}$	Absorption coefficient of divinyl-chlorophyll A at 680 nm	0.0184	$m^2 \text{ (mg dvchl-A)}^{-1}$	Bricaud et al., 2004
$a_{dvchl-B}$	Absorption coefficient of divinyl-chlorophyll B at 680 nm	0.0018	$m^2 \text{ (mg dvchl-B)}^{-1}$	Bricaud et al., 2004
$C_{dvchl-A}$	Amount of divinyl-chlorophyll A	0.0163	$\text{g dvchl-A gDW}^{-1}$	Casey et al., 2016
$C_{dvchl-B}$	Amount of divinyl-chlorophyll B	0.0013	$\text{g dvchl-B gDW}^{-1}$	Casey et al., 2016
d	Average diameter of MED4	0.6	μm	
n'	Imaginary part of the refractive index at 675 nm	0.01377		Stramski et al., 2001
Q^*	Packaging effect at 680 nm	0.945		

(Grossowicz et al., 2017), first by finding the range of possible parameter combinations corresponding to the gross growth rate of 0.5 d^{-1} (**Supplementary Figure 1A**), and secondly by comparing predicted growth and ammonium depletion with the experimental time-series cultivation data (**Supplementary Figures 1B,C**). The estimated parameters were used in the remaining dynamic FBA simulations in COMETS. Finally, to simulate the dynamic storage and consumption of glycogen we applied a multiple objective approach consisting of the following four steps: (1) Maximization of the flux through the non-growth associated maintenance reaction. Note that, this reaction has an upper bound of $1 \text{ mmol gDW}^{-1} \text{ h}^{-1}$ (Casey et al., 2016). In contrast to standard practice, where one uses a lower bound for the non-growth associated maintenance reaction, this method provides a more realistic scenario where the organism continues to consume resources trying to keep up cellular maintenance even at zero growth; (2) Maximization of growth; (3) Maximization of glycogen production (storage); and (4) Parsimonious objective which minimizes the sum of absolute fluxes. To simulate nitrogen-abundant and nitrogen-poor growth conditions, we used the PRO99 medium with standard ($800 \mu\text{Mol}$) and reduced ($100 \mu\text{Mol}$) ammonium concentration, as previously described (Grossowicz et al., 2017). Light availability was modeled as described in Equation 5, with an amplitude of $40 \mu\text{mol Q m}^{-2} \text{ s}^{-1}$ and a period of 24 h. We also incorporated a death rate of 0.1 d^{-1} , similar to previous modeling efforts on *Prochlorococcus* (Grossowicz et al., 2017). All parameter values used in the COMETS simulations are given in **Supplementary Table 5**. All dynamic growth simulations were performed using COMETS v.2.7.4 with the Gurobi 8.1.1 solver, invoked using the associated MATLAB toolbox¹.

Simulating Growth of Knockout Mutants

Simulations of the knockout mutants were performed by constraining the flux to zero for the reactions catalyzed by the enzymes encoded by *glgC* (PMM0769) and *gnd* (PMM0770), respectively. For *glgC*, the reaction is glucose-1-phosphate adenylyltransferase (R00948) and for *gnd* the two reactions are NADP^+ and NAD^+ associated 6-phosphogluconate

dehydrogenases (R01528 and R10221). We then used dynamic FBA in COMETS with PRO99 medium (Moore et al., 2007) with limited ammonium and diel light conditions to simulate growth over 7 days. The growth curves were qualitatively compared with experimental data from Shinde et al. (2020).

RESULTS AND DISCUSSION

Model Curation and Update

Prochlorococcus fixes carbon through photosynthesis during daytime. Fixed carbon that is neither used for cell growth nor stored in the form of glycogen is exuded. Here, we set out to study dynamic changes in the carbon allocation and storage mechanisms in MED4 using a genome-scale metabolic modeling approach. To that end, we first re-curated and updated the available *i*JC586 model (Casey et al., 2016), as described in detail in the “Materials and Methods” section. The update involved the development of a new semi-automatic algorithm (ReFill), which can be broadly applied to other reconstructions (see section Materials and Methods). Concurrently, we introduced a revised mechanism for carbon storage, effectively treating glycogen as an independent component of biomass. This dynamic implementation of glycogen storage, introduced here in dFBA, makes it possible for glycogen to be accumulated and depleted at variable rates (**Figure 1**), aligning with the overflow metabolism hypothesis (Szul et al., 2019; de Groot et al., 2020). Other key modifications induced by the ReFill algorithm and subsequent manual curation (see section Materials and Methods) include the completion of the Entner-Doudoroff (ED) pathway, recently discovered in cyanobacteria (Chen et al., 2016) and proposed as the primary *Prochlorococcus* glucose metabolism pathway under mixotrophic conditions (Billier et al., 2018; Muñoz-Marín et al., 2020). Additional revisions focused on the exudation of fixed carbon products from the cell and included various transports such as pyruvate, fumarate, citrate, ethanol, various nucleotides and hydrogen peroxide as well as metabolites found in both the endo- and exo- metabolome of *Prochlorococcus* (Metabolights study MTBLS567). The end product of our revision, reconstruction *i*SO595, has 595 genes, 802 metabolites and 994 reactions, i.e., 27 genes, 123 metabolites and 196 reactions more than the previous version, *i*JC568 (**Figure 1D**).

¹<https://github.com/segrelab/comets-toolbox>

Carbon Fixation and Storage Are Affected by Nutrient Uptake Rate

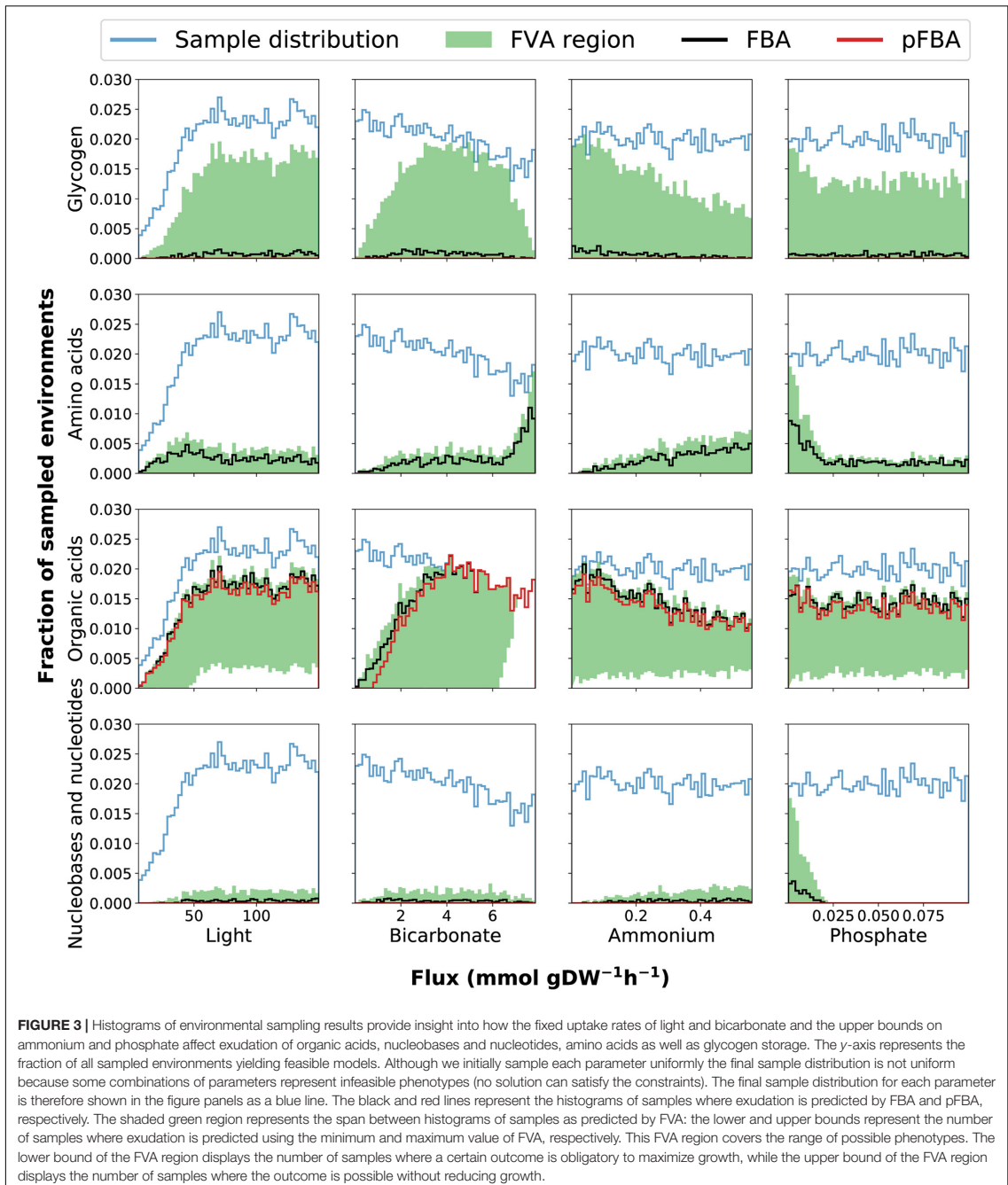
Prochlorococcus thrive in oligotrophic environments (Johnson et al., 2006), where, in surface waters, its growth and carbon fixation rates are usually limited by the abundance of nitrogen, phosphate or iron (Krumhardt et al., 2013; Saito et al., 2014; Szul et al., 2019). Deeper in the water column *Prochlorococcus* growth becomes limited by light (Vaulot et al., 1995). We set out to explore the combined effect of different levels of light and nutrients on carbon fixation, storage and exudation. Similarly to Phenotypic Phase Plane analysis (Edwards et al., 2002), we sought a global perspective of metabolism in this multi-parameter spaces while explicitly taking into account the fact that the inflow of light and bicarbonate may not be easily controllable by the cell, and that *Prochlorococcus* may need to deal with excess amounts of fixed carbon. Thus, in contrast to normal FBA where the uptake of metabolites is constrained by an upper bound, we introduced a ‘push-FBA’ approach (Figure 1A), in which the influx of bicarbonate and light have a fixed imposed value (see section “Materials and Methods” and Supplementary Table 3 for specific values used). This approach attempts to mimic implications of photosynthesis, in which light is the driving force. Once photons are absorbed by the chlorophyll in the photosynthetic reaction centers, most of the energy must be used to produce ATP and reducing power, otherwise it is dissipated in ways that may cause cell damage (Long et al., 1994). We note that this modeling approach over-simplifies the complex process of photosynthesis; for example, we do not account for the dynamics of photoprotective pigments, which allow some of the incident photons to be dissipated as heat. Indeed, the ratio of the photoprotective pigment zeaxanthin to divinyl chlorophyll *a* increases under nitrogen starvation, suggesting that, under these conditions, some of the photon flux may be diverted from the reaction centers (Steglich et al., 2001; Roth-rosenberg et al., 2019). Nevertheless, *Prochlorococcus* undergo photoinhibition at high light intensities (Moore et al., 1995; Mella-Flores et al., 2012), despite the presence of photoprotective pigments and other protection mechanisms such as cyclic electron flow [which is represented in the model (Casey et al., 2016)]. Thus, these mechanisms do not allow the cell to fully control the flux of photons through the photosystem and the resulting fluxes in ATP and reducing power, in a manner that is reflected in the push-FBA approach. This subtle difference in applied constraints has major effects on model predictions. While flux rearrangement is usually viewed as a consequence of environmental nutrient limitations, the results of this analysis show that a substantial rewiring of fluxes is caused by this imposed excess of fixed carbon as well.

To understand how different combinations of environmental parameters (availability of nitrogen, phosphate, light and bicarbonate) affect the way *Prochlorococcus* can manage its carbon budget, we implemented FBA under 10,000 randomly sampled growth environments. Overall, this sampling analysis demonstrated that the exudation of organic acids, amino-acids, and nucleobases/nucleosides, as well as the extent of glycogen storage, are strongly modulated by environmental factors (Figure 3). To observe the full range of possible optimal

solutions per sample, we implemented and compared different flux balance analysis methods, including flux variability analysis (FVA) and parsimonious FBA (pFBA). These two methods provide complementary insight: FVA estimates the range of possible values for the flux of each reaction at the optimum, providing insight into the structure of the phenotypic space at maximal growth rate. In contrast, pFBA, by minimizing the sum of fluxes at optimality, generates flux predictions less likely to involve unrealistic loops, and thus potentially provides predictions closer to experimental values (Lewis et al., 2010). Together, these two FBA methods help analyze the solutions of our high-dimensionality dataset.

Our predictions simulate the metabolic effects and variability in glycogen production modulated by environmental constraints (Figure 3). Glycogen production was observed only above light levels of $50 \text{ mmol gDW}^{-1} \text{ h}^{-1}$ (corresponding to $7.5 \text{ micromole quanta m}^{-2} \text{ sec}^{-1}$), and decreased as ammonium and phosphate concentrations increase. These observations do not contradict previous evidence showing increased glycogen accumulation in faster growing cyanobacteria (Zavřel et al., 2019), rather they align with previous studies finding that glycogen storage is enhanced in nutrient-limiting conditions (Monshupanee and Incharoensakdi, 2014; Szul et al., 2019). Interestingly, FVA consistently predicted the glycogen production range minimal value to be zero across all samples. This implies that glycogen storage is possible, but not necessary to achieve optimal growth in the feasible solution space. This was also the case in the more stringent pFBA analysis, indicating that while metabolism may be a strong modulator of glycogen metabolism, more types of regulation, not accounted for in FBA, are involved. One example of such regulation may be allosteric regulation of ADP-glucose pyrophosphorylase by 3-phosphoglycerate (Iglesias et al., 1991), possibly in combination with redox regulation (Díaz-Troya et al., 2014). Specific regulation aimed at tuning up glycogen storage may also occur at the transcriptional level, e.g., by multiple transcription factors previously suggested to be involved in the regulation of glycogen metabolism in fluctuating environments (Luan et al., 2019).

The range of possible rates of glycogen production (through FVA) displays a bell-shaped bicarbonate-dependent distribution, indicating low storage of glycogen (zero flux) under both low and high uptake rates of bicarbonate. When bicarbonate uptake rates are low, all available carbon is diverted into growth. The reduced glycogen storage at high bicarbonate uptake, when RuBisCO is saturated, seems to be caused by the increased ATP demand associated with the conversion of bicarbonate to exudation-products, since the onset and rate of change of this trade-off is modulated by the ATP availability, as demonstrated by phenotypic phase planes analysis (Supplementary Figure 2). This agrees with recent work suggesting that *Prochlorococcus* use available ATP to drive pathways to saturation by shifting reaction directions toward favoring dephosphorylation of ATP to ADP, disrupting the cellular ATP/ADP ratio and increasing the metabolic rate of the cell by pushing forward ATP consuming reactions, until it is restored. Together with organic carbon exudation this strategy allows for growth in lower nutrient concentrations (Braakman, 2019).



We next sought to explore the effect of combinations of key nutrients on storage and exudation patterns in our sampling spaces. To that end, we visualized the data using t-SNE clustering

(Figure 4A). To explore the strongest trends, we chose to employ a high stringency approach and use only our set of pFBA results in this context. Due to the nature of pFBA,

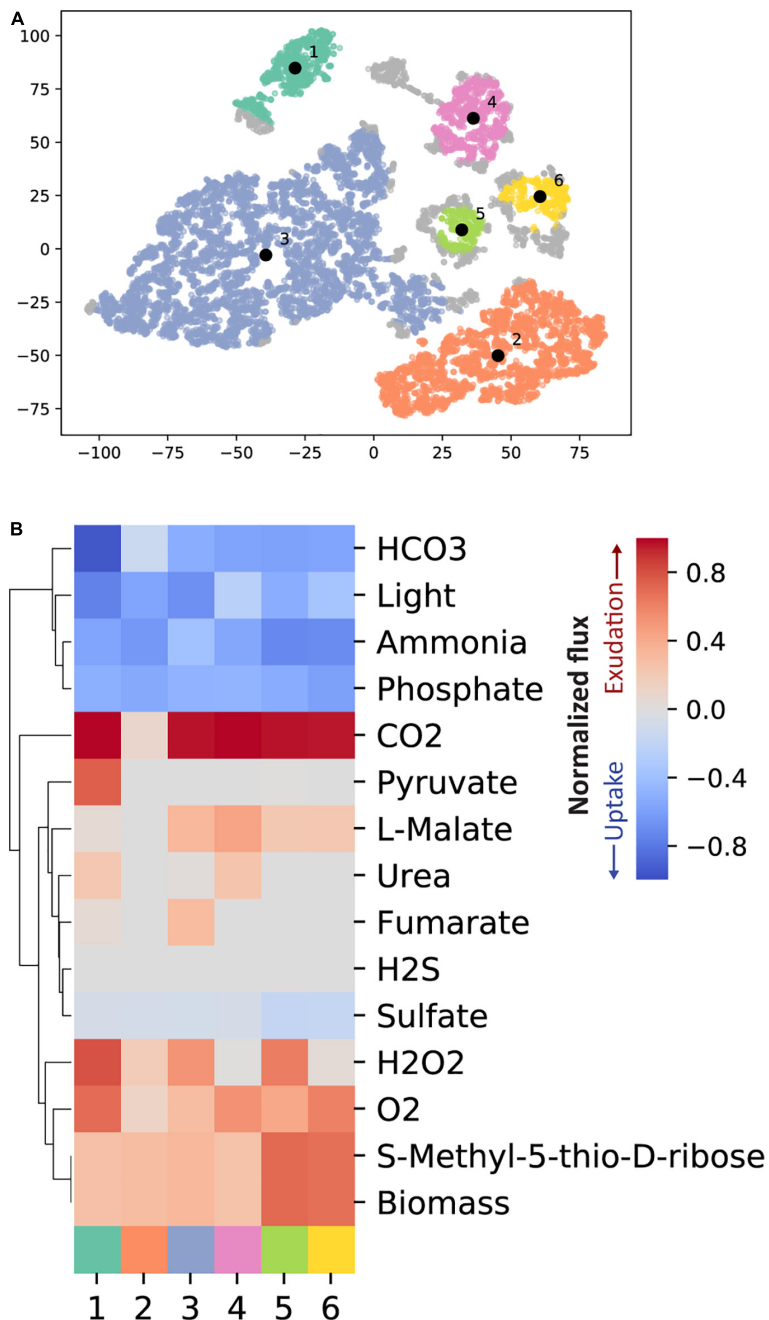


FIGURE 4 | T-SNE clustering identifies typical phenotypes from the pFBA results from the random samples. **(A)** The random samples are reduced into two dimensions with t-SNE. We have subsequently used HDBSCAN to cluster the data. HDBSCAN identified six disjoint clusters which represent different phenotypes. **(B)** For each of the six clusters the mean uptake or exudation across all samples within the respective cluster is shown. Only exchange reactions with an absolute flux above $1e-3 \text{ mmol gDW}^{-1} \text{ h}^{-1}$ in any of the random samples are included.

any exudation observed in this analysis could not be easily removed without imposing a cost on growth. We observed 6 typical phenotypes (clusters) rising out of the sampling spaces (Figure 4). These 6 phenotypes are characterized by subtle differences in combinations of environmental parameters, yielding significantly different exudation patterns. Generally, we observed the highest biomass value in phenotype 5, and the lowest in phenotype 4. All key nutrient uptake rates were highly variable (ranging from 33 to 44% variability). Phenotype 1 is characterized by high light, bicarbonate, a maximum RuBisCO flux (indicating maximal photosynthesis rate) but low nitrogen uptake. Additionally, we observed high exudation of pyruvate coming from the pentose phosphate and Entner–Doudoroff pathways. Both are alternative routes coming out of carbon fixation (Waldbauer et al., 2012; Chen et al., 2016). Together with a low biomass value, this phenotype might indicate a scenario of exudation due to overflow metabolism.

The two largest clusters (numbers 2 and 3, Supplementary Figure 3), tie together high and low light, carbon and nitrogen uptake rates, and different exudation patterns. Interestingly, phenotype 3 (high light) showed exudation of fumarate and malate while phenotype 2 (low light) did not. Recent work suggested that, in high light conditions, fumarate is generated through oxaloacetate and malate creating a broken acyclic form of the TCA cycle, while in the dark, fluxes are diverted into forming the cyclic form of it. This low light form of the TCA cycle is then active and works toward energy generation (Xiong et al., 2017). Similarly, we observed two forms of the TCA cycle in the high and low light phenotypes (2 and 3, respectively) with a difference in the direction of one reaction (KEGG R00342, Supplementary Figure 3). Phenotype 2, describing low-light conditions, showed the L-Malate/oxaloacetate balance to shift in favor of oxaloacetate, completing the route toward 2-Oxoglutarate, a key metabolite known to act as a starvation signal and modulator of the C/N balance in cyanobacteria (Dominguez-Martín et al., 2018; Zhang et al., 2018), and subsequently into energy generation. On the other hand, Phenotype 3, describing high light conditions, showed the L-Malate/oxaloacetate balance to shift in favor of L-Malate and away from the formation of 2-oxoglutarate. In both phenotypes fumarate is converted to L-Malate. While in Phenotype 2 it is fed into a semi-cyclic form of the TCA cycle, fumarate is partly exuded and partly converted to L-malate in phenotype 3, in agreement with overflow metabolism.

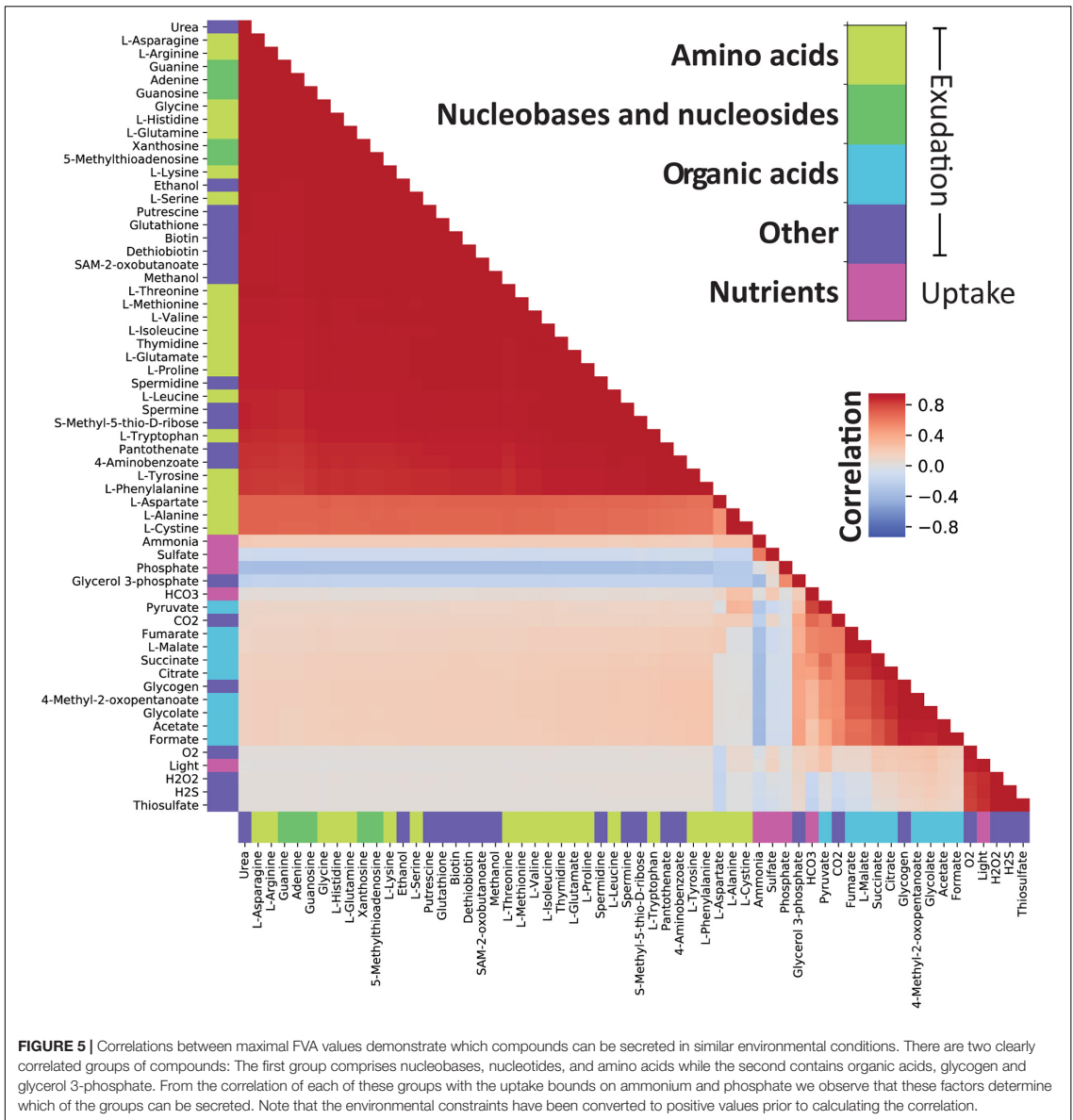
We observed a similar TCA cycle flux distribution in phenotype 4 as in phenotype 3, leading to high exudation of L-Malate. Interestingly, Phenotype 1 and 4 are comparable in all key nutrients except light (High in phenotype 1 and low in phenotype 4). As a result of an in-depth flux distribution analysis, we observed a reaction direction change in UDP-glucose:NAD⁺ 6-oxidoreductase [R00286, EC 1.1.1.22, PMM1261] between the two phenotypes. In phenotype 4 this reaction shifted toward the creation of UDP-glucose, a precursor for the production of glycogen (due to the high stringency of this analysis we did not observe the direct formation of glycogen). In phenotype 1, this reaction favored the formation of UDP-glucuronate which in turn was diverted into the formation of amino sugars. These phenotypes may correlate to the 12:00 (phenotype 1) and 16:00

(phenotype 4) scenarios described in Szul et al. (2019). Finally, Phenotypes 5 and 6 may represent a high-light nutrient-rich environment resulting in a high biomass value.

Nutrient Uptake Rates Modulate Exudation of Organic Compounds

The use of genome-scale metabolic models captures a comprehensive picture of the metabolic processes taking place in the cell, including those that lead to metabolite exudation. From the random sampling of environmental conditions, we identified conditions in which organic acids must be exuded. This was noticeable by a non-zero lower bound of the FVA region (Figure 3). Interestingly, organic acids were more likely to be exuded when the growth became limited by phosphate or nitrogen. Since *Prochlorococcus* is known to thrive in oligotrophic ocean gyres where nitrogen or phosphate is limited (Partensky et al., 1999; Flombaum et al., 2013), this represents a likely natural phenotype, and as such, supports previous findings (Bertilsson et al., 2005; Szul et al., 2019). Costly metabolites, essential for cell survival and growth, such as amino acids, nucleobases and nucleotides, tend to be exuded in nitrogen and carbon rich conditions and might be a result of overflow metabolism (Cano et al., 2018; Pacheco et al., 2018). To explore this phenomenon in further detail, we looked into exudation patterns of specific metabolites as a function of key nutrient limitations (Figure 5). Of the environmental factors, the uptake of nitrogen (ammonium) is a decisive factor differentiating between exudation of organic acids or amino acids. While it is positively correlated with the exudation of nitrogen-rich compounds such as amino acids, it is negatively correlated with exudation of organic acids and glycogen. Additionally, glycogen formation is positively correlated with the exudation of malate, citrate, fumarate, and succinate, which are most of the TCA cycle constituents. This is in line with previous findings suggesting the re-direction of carbon metabolism toward the formation of macromolecules (including glycogen) in nitrogen limiting conditions (Forchhammer and Selim, 2019; Szul et al., 2019). Thus, our reconstruction captured known possible aspects of the carbon/nitrogen balance in *Prochlorococcus*.

Finally, we observed a general pattern of strong positive correlations between amino acids, nucleobases, nucleosides, as well as a range of other compounds. In an interesting deviation from this general pattern, L-aspartate showed a decreased correlation with other exudates. L-aspartate, together with its role in protein nucleic acid biosynthesis, can serve as a precursor for nitrogen storage metabolites such as polyamines (Szul et al., 2019). Indeed, we observed a slightly stronger correlation between L-aspartate and the uptake of nitrogen compared to other amino acids. Finally, In contrast to other amino acids, L-aspartate is negatively correlated with light uptake and hydrogen peroxide exudation. Hydrogen peroxide is produced from L-aspartate and oxygen by L-aspartate oxidase [R00481, EC 1.4.3.16, PMM0100]. L-amino acid oxidases have been previously described in cyanobacteria and have been related to the use of amino acids as carbon sources (Campillo-Brocal et al., 2015). The production of hydrogen peroxide is also strongly correlated with



light, a result consistent with the expectation that reactive oxygen species are created during photosynthesis.

Dynamic Allocation of Carbon Storage

Nutrient and light limitations are well-known modulators of carbon storage in *Prochlorococcus* (Zinser et al., 2009; Szul et al., 2019). Recent work has suggested the storage of carbon to be one of the major metabolic tasks during the day-night cycle (Cano et al., 2018; Szul et al., 2019; Shinde et al., 2020).

To explore time-modulated trade-offs and trends related to carbon storage, we performed *in silico* dynamic FBA diel-cycle simulations using the Computational of Microbial Ecosystems in Time and Space (COMETS) platform (Harcombe et al., 2014; Dukovski et al., 2020). COMETS is a population-based dynamic FBA implementation that can simulate growth of millions of cells, but it is important to note that the framework assumes continuous growth on a mesoscopic scale and does therefore not explicitly account for individual cells nor regulated cell

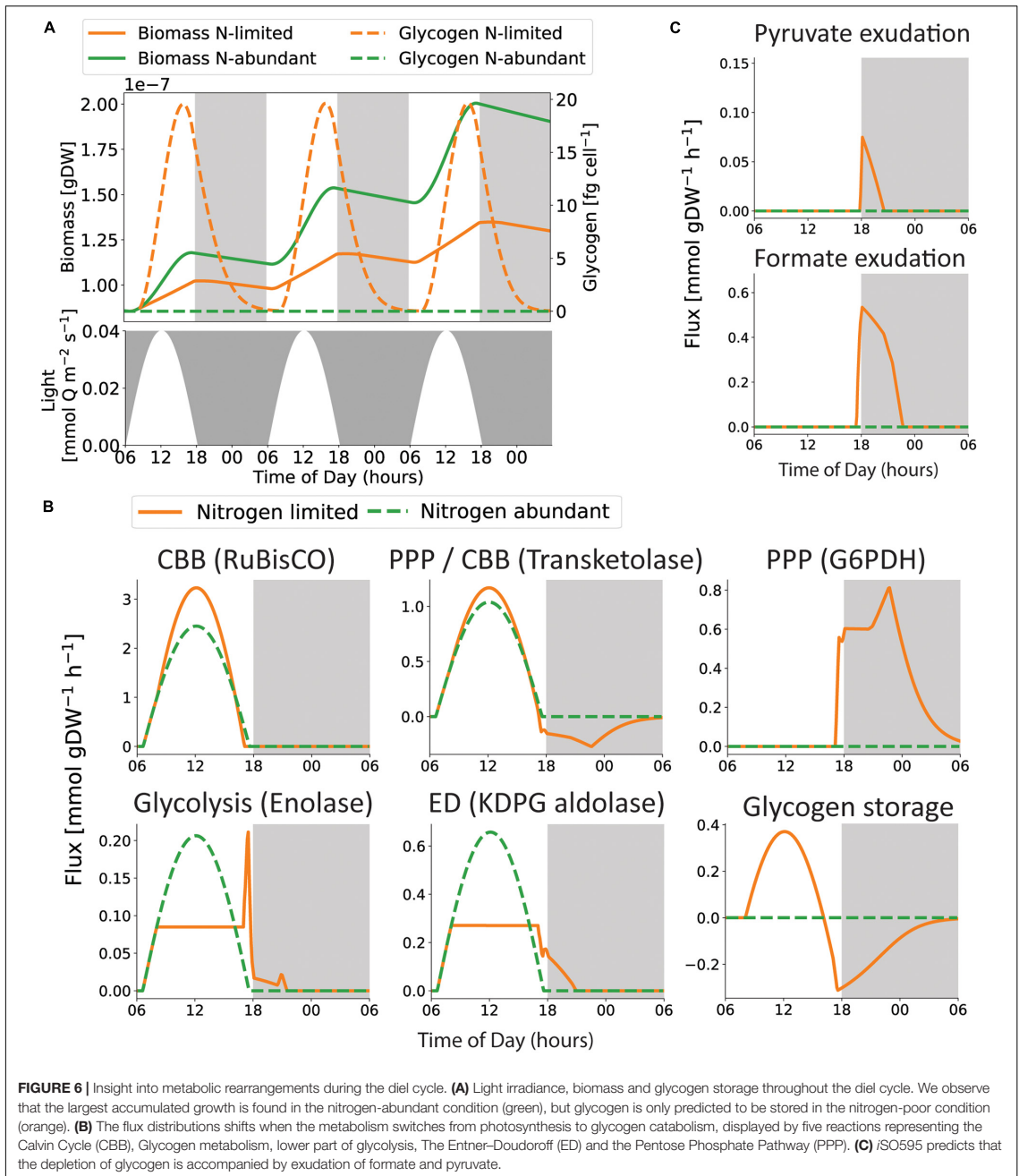
cycle events such as cell division. COMETS relies on uptake flux kinetic information such as K_m and V_{max} to simulate the spatial growth and exudation patterns of microbes in a simulated discretized time course. To improve the accuracy and biological relevance of our simulations we used kinetic constants either obtained from experimental measurements reported in the literature (Krumhardt et al., 2013; Hopkinson et al., 2014) (Supplementary Table 3) or from fitting model simulations to measured growth and depletion of ammonium rates (Grossowicz et al., 2017). We found K_m and V_{max} values of 0.39 mM and 0.9 mmol gDW⁻¹ h⁻¹ for the uptake of ammonium to best fit the experimental data (Grossowicz et al., 2017) (Supplementary Figure 1). Surprisingly, the estimated K_m value is 3 orders of magnitude larger than previous estimates (Marañón et al., 2013). This deviation might occur due to several reasons. First, our estimates are based on the assumption that growth is indeed limited by the availability of ammonium and that *Prochlorococcus* operates at a metabolic state close to optimal growth. Other limiting factors or non-optimal growth may lead to incorrect estimates. Nevertheless, it is challenging to fit K_m values accurately from batch cultivation data, as this parameter only becomes dominant in the short time-period immediately prior to nutrient depletion. Furthermore, the accuracy of the fitted K_m value can suffer from the rather high uncertainty in the measured ammonium concentrations, although not more than 2 orders of magnitude (Supplementary Figure 1). Finally, we raise the possibility that *Prochlorococcus* may possess several ammonium transporters with different affinity as previously observed in marine eukaryotic phytoplankton (McDonald et al., 2010) and cyanobacteria (Kashyap and Singh, 1985). To account for this uncertainty we assessed the sensitivity of our dFBA simulations to variation in the value of K_m , in combination with variation in the maximum uptake rate of ammonium (V_{max}), ammonium concentration and light intensity (Supplementary Figure 4). The parameters that dictate light absorption (Table 1) affect the number of available photons, so that by including a large span of light intensities in our sensitive analysis, we also cover their associated uncertainty. We find that ammonium concentration, kinetic coefficients for ammonium uptake and the availability of photons combined have a considerable impact on whether carbon is stored during daytime in our dynamic FBA simulations, underpinning the importance of accurate and context specific values for these parameters. This echoes the well-known modulation of carbon storage by nutrient and light limitations (Zinser et al., 2009; Szul et al., 2019). We note that, despite the potentially large impact of *Prochlorococcus* on marine nitrogen budgets, to the best of our knowledge there are currently no direct experimental measurements of the kinetics (K_m , V_{max}) of nitrogen uptake by *Prochlorococcus*.

Since the tight coupling between carbon and nitrogen metabolism in cyanobacteria is known to influence carbon allocation and storage (Zhang et al., 2018; Szul et al., 2019), it was chosen as a case study. As such, we focused in more detail on the dynamic changes in metabolism in nitrogen-abundant and nitrogen-poor media, as previously defined (Grossowicz et al., 2017). Specifically, we set out to explore glycogen production and consumption with COMETS in these conditions (Figure 6).

We did not observe glycogen storage in nitrogen-abundant simulations, and therefore no growth nor cellular maintenance during nighttime. One explanation for this may arise from the limitations of the platform. First, the simulations performed in this work were performed in a modeling framework based on linear programming with ordered multi-objective optimization: (1) cellular maintenance; (2) growth; (3) glycogen storage. Thus, glycogen was only stored when there were excess energy and carbon available, which occurred when growth was nitrogen limited. Although some observer bias was introduced by assuming that *Prochlorococcus* is striving toward these cellular objectives, in this order, we found a reasonable conceptual alignment with previous work showing that bacterial metabolism balances a trade-off between maximal growth and the ability to adapt to changing conditions (Schuetz et al., 2012). However, we do note that one might obtain more nuanced results by taking into account suboptimal solutions (Segrè et al., 2002; Fischer and Sauer, 2005; Wintermute et al., 2013), and that real phenotypes may be in the continuum between the two extremes found here. Another limitation that might affect glycogen storage is the lack of regulatory mechanisms not usually accounted for in this version of dynamic FBA (Mahadevan et al., 2002). The addition of regulatory layers or more specifically tailored objective functions, such as global optimization over the entire diel cycle (Reimers et al., 2017), could lead to smaller but non-zero generation of glycogen also during nitrogen-rich conditions.

In agreement with previous work (Szul et al., 2019), under nitrogen-limiting conditions, glycogen accumulates throughout the day and is subsequently used to support respiration and growth during the night (Figure 6A). However, the predicted glycogen storage is simulated as not sufficient to support neither growth nor cellular maintenance throughout the night. This may contribute to the increased death rate during night time (Zinser et al., 2009; Ribalet et al., 2015). However, the rate of glycogen depletion is strongly affected by the associated kinetic parameters (Supplementary Figure 5), emphasizing the value of accurate kinetic coefficients for GlgP, the main contributing factor to glycogen catabolism in bacteria (Dauvillée et al., 2005; Alonso-Casajús et al., 2006; Fu and Xu, 2006), in future work. Furthermore, the rate of glycogen depletion might be modulated by transcriptional regulation. Previous work suggested that glycogen storages are not sustained beyond dawn, because the genes responsible for glycogen degradation are depleted during the first 5 h of darkness (Biller et al., 2018). Interestingly, the model predicts consumption of glycogen during dusk to increase growth when photosynthesis is declining (Figure 6), closely resembling observations in *Synechococcus*, in particular for the $\Delta kaiC$ mutant with a dysfunctional circadian clock (Diamond et al., 2015). The closer resemblance of the dysfunctional circadian clock phenotype might be a result from the limitations of the applied modeling framework that does not include regulatory mechanisms.

The switch from photosynthesis at daytime to glycogen consumption at nighttime is reflected in the metabolic shifts observed in key pathways (Figure 6B). Interestingly, we observed higher fluxes through the Calvin cycle in nitrogen-poor conditions. This difference may be caused by the increased



ATP demand necessary to support higher growth rates in nitrogen-abundant conditions. Additionally, our simulations predicted that the use of the Entner-Doudoroff pathway during photosynthesis creates precursor metabolites for growth during

light hours, and a shift to the Pentose Phosphate Pathway (PPP) during nighttime. This trend might occur as an alternative for generating NADPH (**Supplementary Figure 6**). Upregulation of the PPP enzymes during dusk and the first half of the night time

was also observed in the proteome of *Prochlorococcus* (Waldbauer et al., 2012). Several enzymatic transformations participate in both the Calvin cycle and the PPP, although in opposite directions (Waldbauer et al., 2012). These transformations were captured in our simulations, specifically as demonstrated by transketolase (Figure 6B). Additionally, the consumption of glycogen during nighttime might lead to exudation of pyruvate and formate (Figure 6C). This prediction is supported by recent observations; formate is exuded during both nutrient-replete and phosphate-limited growth in *Prochlorococcus* strains MED4 and MIT9312 under constant light (Bertilsson et al., 2005), as well as when phosphonates are metabolized in *Prochlorococcus* strain MIT9301 (Sosa et al., 2019). Thus, *Prochlorococcus* are potential formate sources for heterotrophs. However, degradation of phosphonates yields formate as an immediate byproduct, and the current modeling framework is not suited to evaluate whether an equally high amount of intracellular formate is feasible during glycogen degradation, as intracellular metabolite concentrations are not readily represented in dFBA. Pyruvate exudation in *Prochlorococcus* is indicated from previous co-cultivations with SAR11 (Becker and Hogle, 2019), and from upregulation of genes encoding pyruvate kinase and a pyruvate efflux transporter during extended darkness (Biller et al., 2018). Furthermore, pyruvate is exuded when fixed carbon is consumed in the closely related strains *S. elongatus* PCC 7942 and *S. sp.* PCC 6803 (Carrieri et al., 2012; Benson et al., 2016).

The shift from photosynthesis and carbon fixation to glycogen catabolism is also associated with a switch in production and consumption of energetic cofactors (Supplementary Figure 6). Generation of ATP is performed concomitantly by ATP synthase in both the thylakoid membrane and the periplasmic membrane during photosynthesis. The periplasmic ATP synthase is first driven by reduced cofactors (NADPH) generated by the electron transport chain in the light-dependent part of photosynthesis (Supplementary Figure 6). ATP is consumed by two separate processes: growth- and maintenance-associated reactions reach a threshold once growth is limited by the nitrogen abundance, while the recycling of precursors for the Calvin cycle follows the shape of light absorption throughout the day. In agreement with previous work (Park and Choi, 2017), our model predicted higher rates of NADPH production than NADH.

Next, we explored the ability of our model to dynamically capture biologically relevant phenotypes by performing dynamic FBA simulations of knock-out mutants in *Prochlorococcus*, focusing on two gene deletions disrupting different parts of glycogen metabolism. $\Delta glgC$ breaks synthesis of ADP-glucose and thus the storage of glycogen and Δgnd , knocking out 6-phosphogluconate dehydrogenase, a key reaction in the Pentose Phosphate pathway found to fuel the Calvin cycle with precursor metabolites during the onset of photosynthesis (Shinde et al., 2020). Our dynamic FBA simulations in COMETS (Supplementary Figure 7) showed similar growth between Δgnd and the wild type and slightly lower growth for $\Delta glgC$. We set out to compare these observations with available experimental data. Since genetic tools for the modification of *Prochlorococcus* are still lacking (Laurenceau et al., 2020), we chose data from the closely related cyanobacteria *Synechococcus* as recent work

described the impact of $\Delta glgC$ and Δgnd on its growth during diel cycles (Shinde et al., 2020). Indeed, we found very good agreement between measured and predicted growth for both the wild-type and $\Delta glgC$ mutant where glycogen storage is disrupted (Shinde et al., 2020) (Supplementary Figure 7). One of the notable limitations of dynamic FBA is the ability to quantify intermediates and precursor pools that might drive the initiation of a pathway. This comes mainly from the assumption of a quasi steady-state of intracellular metabolite pools at each time point. Although the comparison is strictly qualitative and concerns strains with known differences (Mary et al., 2004), these findings demonstrated the ability of our reconstruction to capture metabolic trends in response to genetic perturbations, indicating that iSO595 will be a valuable tool in future research of *Prochlorococcus*. Overall, our dynamic simulations display biological and physiological behaviors that are consistent with expectations, and at the same time provide valuable insight into the putative internal metabolic processes that might modulate the *Prochlorococcus* growth under environmental and genome-induced constraints.

CONCLUSION

Our study provides a detailed systematic view of the underlying metabolic trends modulating carbon storage and exudation in *Prochlorococcus*. *Prochlorococcus* is known to interact with other bacteria in its surroundings (Sher et al., 2011; Aharonovich and Sher, 2016; Biller et al., 2018; Hennon et al., 2018). It is currently impossible to predict the fluxes of organic matter (or of the myriad metabolites comprising it, such as amino acids, sugars, and organic acids) between phytoplankton and bacteria. Yet, quantifying such fluxes and predicting them from genomic surveys, as shown here, serves a number of roles: (1) It can provide experimentally testable and mechanistic hypotheses on inter-microbial exchanges and competition, (2) It has the potential to increase knowledge about the specific metabolites that may mediate these interactions; and (3) It would enable the construction of improved models of biogeochemical cycles which consider the diverse and powerful metabolic capabilities of the ocean microbiome.

Genome-scale metabolic-network reconstructions are powerful tools, but not without limitations. Mainly, the predictive accuracy rests on the quality and completeness of the metabolic network. The construction and curation of these metabolic networks depend heavily on data availability and annotation accuracy, which may be scarce for less studied organisms. Several methods have been developed to fill the gaps of incomplete network reconstructions. For example, FastGapFill incorporates missing knowledge from universal, non-organism specific data (Thiele et al., 2014), ModelSEED fills gaps through the use of thermodynamic parameters and FBA simulations to achieve minimal growth (Henry et al., 2010), and MENECO uses a topology graph based approach to look for minimal sets of metabolic reactions that support growth and the producibility of target metabolites (Prigent et al., 2017). In this work we used a novel semi-automated gap-filling method (ReFill) to

increase existing knowledge in the reconstruction by up to 25%. In contrast to other standard gap filling approaches, ReFill has the specific capability to add individual reactions through a recursive algorithm that guarantees complete connectivity to the existing network, incorporating the maximal possible amount of validated, organism specific metabolic annotations. However, this approach employs high stringency and thus adds limited amounts of knowledge. Considering that *Prochlorococcus* strains have some of the smallest known genomes among free-living organisms, a 25% increase in knowledge serves as a significant improvement in the predictive capacity of the model. However, reconstruction of high-quality genome-scale metabolic models is an iterative process, where new data, knowledge, and scope create opportunities for further model improvement. One example of this possibility is the CO₂ concentrating mechanisms in *Prochlorococcus*. This mechanism is known to be sustained by proton and ion gradients across the cell membrane at an energetic cost (Hopkinson et al., 2014; Burnap et al., 2015). However, the comprehensive knowledge and annotation of ion transporters necessary to model this mechanism are lacking, and are therefore not included in iSO595. With the advancement of data collection and annotation tools, together with the use of ReFill or similar algorithms, metabolic knowledge can be added to such reconstructions, improving their predictive abilities and mimicry of biological and physiological processes.

Other limitations of static and dynamic FBA simulations include the inability to represent metabolite concentrations and the lack of regulatory effects. Furthermore, since COMETS, like most other implementations of dFBA, simulates millions of asynchronously growing and dividing cells on the mesoscopic scale, cell cycle processes are not readily incorporated into this framework. Thus, future extensions to this work include the implementation of cell division in *Prochlorococcus*, known to occur in the afternoon (Vaulot et al., 1995). Another improvement would be an accurate representation of the costs associated with light damage and the production of protective pigments required to combat excessive light absorption. This could potentially be accounted for by extending the current *Prochlorococcus* GEM to a framework that includes macromolecular allocation, such as Resource Balance Analysis (Goelzer et al., 2011), conditional FBA (Rügen et al., 2015) or models of metabolism and macromolecular expression (ME models) (Thiele et al., 2012). Along these lines, one could relate mortality with an inability to maintain basic cellular functions, rather than a fixed death rate. However, the relationship between cell mortality and metabolism is not well constrained, and its representation in dFBA models is currently rudimentary. Future work is needed to better understand mortality and represent it in models of cell metabolism, ecosystems and biogeochemistry. Finally, our findings contribute to a growing body of work on the underlying metabolic mechanisms modulating the metabolic success of *Prochlorococcus*. The approaches shown here provide systematic insights corroborated in recent and well-known works and provide strong foundations for future studies of *Prochlorococcus* metabolism with particular interest in its interaction with other microorganisms and the effects of these on community composition and larger biogeochemical cycles.

DATA AVAILABILITY STATEMENT

Publicly available datasets were analyzed in this study. This data can be found here: KEGG (through Python API): <https://www.genome.jp/kegg/>, TransportDB: <http://www.membranetransport.org/transportDB2/index.html>, Metabolights: <https://www.ebi.ac.uk/metabolights/MTBLS567>, and former *Prochlorococcus* model: <https://msystems.asm.org/content/1/6/e00065-16>.

AUTHOR CONTRIBUTIONS

DSh and DSe designed the study. SO, SS, and DSe developed the computational models and performed the computational analyses, with input from DSh. SO and SS wrote a first version of the manuscript. DSe, DSh, and EA oversaw the project and contributed to the final version of this manuscript. All authors have read and approved the final version of the manuscript.

FUNDING

This work was supported by the Human Frontiers Science Program (grant RGP0020/2016) and the National Science Foundation (NSFOCE-BSF 1635070) to DSe and DSh. DSe also acknowledges support by the National Science Foundation (grant 1457695), the Directorates for Biological Sciences and Geosciences at the National Science Foundation and NASA (agreement nos. 80NSSC17K0295, 80NSSC17K0296 and 1724150) issued through the Astrobiology Program of the Science Mission Directorate, and the Boston University Interdisciplinary Biomedical Research Office. SS was funded by SINTEF, the Norwegian graduate research school in bioinformatics, biostatistics and systems biology (NORBIS) and by the INBioPharm project of the Centre for Digital Life Norway (Research Council of Norway grant no. 248885).

ACKNOWLEDGMENTS

We are grateful to members of the Segrè lab for constructive feedback on the manuscript, and to members of the labs of DS, Hans-Peter Grossart, and Maren Voss for helpful and fun discussions on marine microbes.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: <https://www.frontiersin.org/articles/10.3389/fgene.2021.586293/full#supplementary-material>

Supplementary Figure 1 | Estimation of kinetic parameters for the uptake of ammonium in *Prochlorococcus*. **(A)** All combinations of K_m and V_{max} along the red trajectory matches the observed gross growth rate of 0.5 d^{-1} (Grossowicz et al., 2017). However, when we compare the dynamics of cell density **(B)** and ammonium concentration **(C)** we find that the best overall prediction is achieved

using $K_m = 0.39 \text{ mM}$ and $V_{\max} = 0.9 \text{ mmol gDW}^{-1} \text{ h}^{-1}$ (marked by an orange dot in **A**).

Supplementary Figure 2 | ATP availability influence modulates the trade-off between glycogen storage and growth. Phenotypic phase planes (Edwards et al., 2002) illustrate the combined effect of glycogen storage and bicarbonate uptake on the maximal growth rate. Compared to the base model (**A**), we observe how that the trade-off is strongly affected by modulated ATP availability, either from an artificial reaction providing extra ATP (**B**) or by increasing (**C**) or decreasing (**D**) the amount of available light. Increasing ATP allows more glycogen storage without reducing the growth rate.

Supplementary Figure 3 | TCA cycle flux diagram differences between the most common phenotypes. Flux diagrams of the TCA cycle in the most common phenotypes 2 (colored orange) and 3 (colored blue). Reactions are denoted by KEGG reaction ids. Reaction colors correspond to cluster colors presented in **Figure 4**.

Supplementary Figure 4 | Sensitivity analysis of dFBA simulations to variability in ammonium concentration, kinetic coefficients of ammonium uptake and light irradiance. (**A**) These phase diagrams display which combinations of K_m and V_{\max} , describing the uptake of ammonium, that leads to glycogen accumulation (red area) at peak irradiance based on the light amplitude and ammonium concentration used to simulate nitrogen-abundant (left) and nitrogen-poor (right) conditions in **Figure 6**. In the orange area no glycogen accumulation is predicted as the growth is limited by the available light, rather than nitrogen. (**B**) These panels display similar phase diagrams as in (**A**), but for different amplitudes of light irradiance, represented by black curves, and for different ammonium concentrations (indicated on top of each panel). The number written on each black curve represents the light irradiance amplitude in $\mu\text{mol Q m}^{-2} \text{ s}^{-1}$. The ammonium concentration in the top left panel is equal to the concentration in our simulated nitrogen-poor conditions, and thus, the $40.0 \mu\text{mol Q m}^{-2} \text{ s}^{-1}$ line in this panel is identical to the boundary between the two phases in the right panel in (**A**). The range of ammonium concentrations is chosen so that it covers both our simulated environment and the ammonium concentration in oligotrophic oceans. The blue, green, and orange points display the combinations of K_m and V_{\max} used/provided in this work and previous publications, respectively.

Supplementary Figure 5 | Sensitivity analysis of dFBA simulation to different parametrization of glycogen consumption. All panels display results obtained from dFBA simulations in COMETS with different combinations of K_m and V_{\max} , describing the consumption of intracellular glycogen, with line colors and corresponding values as shown in the bottom right corner color matrix. The purple color corresponds to the values used to run the simulations shown in **Figure 6**. (**A**) Growth curves. (**B**) Accumulated glycogen per gram dry weight of biomass. (**C**) Predicted reaction fluxes for the same 8 reactions as shown in **Figure 6**.

REFERENCES

- Aharonovich, D., and Sher, D. (2016). Transcriptional response of *Prochlorococcus* to co-culture with a marine *Alteromonas*: differences between strains and the involvement of putative infochemicals. *ISME J.* 10, 2892–2906. doi: 10.1038/ismej.2016.70
- Alonso-Casajús, N., Dauvillée, D., Viale, A. M., Muñoz, F. J., Baroja-Fernández, E., Morán-Zorzano, M. T., et al. (2006). Glycogen phosphorylase, the product of the *glgP* gene, catalyzes glycogen breakdown by removing glucose units from the nonreducing ends in *Escherichia coli*. *J. Bacteriol.* 188, 5266–5272. doi: 10.1128/JB.01566-05
- Altschul, S., Gish, W., Miller, W., Myers, E., and Lipman, D. (1990). Basic local alignment search tool. *J. Mol. Biol.* 215, 403–410. doi: 10.1016/S0022-2836(05)80360-2
- Amin, S. A., Parker, M. S., and Armbrust, E. V. (2012). Interactions between diatoms and bacteria. *Microbiol. Mol. Biol. Rev.* 76, 667–684. doi: 10.1128/MMBR.00007-12
- Arrieta, J. M., Mayol, E., Hansman, R. L., and Herndl, G. J. (2015). Dilution limits dissolved organic carbon utilization in the deep ocean. *Science* 348, 331–334.

Supplementary Figure 6 | The transition from daytime to nighttime is associated with a drastic change in the production and consumption of the energy-carrying cofactors. The figure panels show the major sources (left) and drains (right) of the cofactors ATP, NADPH and NADH. The legend shows the reaction IDs used in iSO595. (**A**) ATP is produced by both the thylakoid (R00086th) and the periplasmic (R00086p) ATP synthase during daytime, but mostly by the periplasmic ATP synthase (respiration) during nighttime. ATP is consumed by reactions associated with growth (BIOMASS and BProtein), cellular maintenance (Maintenance) and storage of glycogen (R00948) in addition to reactions recycling precursors for the Calvin cycle (R01512 and R01523) and acetyl-CoA carboxylase (R00742). (**B**) NADPH is produced by ferredoxin reductase (fdr) during daytime and by the pentose phosphate pathway (R02736 and R01528) during nighttime. The NADPH is either used to drive the proton gradient across the periplasmic membrane (NADPHDhp) or in Gluconeogenesis (R1063) to refuel the Calvin cycle during photosynthesis. (**C**) NADH production is correlated with the growth rate and dominated by pyruvate dehydrogenase (R00209) during daytime and the glycine cleavage system (R01221) during nighttime. NADH is consumed by 6-phosphogluconate dehydrogenase (reverse, R10221) during daytime, NADH transhydrogenase (R00112) solely during dusk and concomitant with methylenetetrahydrofolate reductase (R07168) during nighttime.

Supplementary Figure 7 | Predicted growth curves show good agreement in a qualitative comparison with experimental growth of *Synechococcus*. To compare growth data we have overlaid growth curves predicted for the wild-type, the ΔgigC -mutant and the Δgnd -mutant of *Prochlorococcus* with experimental OD measurements of *Synechococcus elongatus* PCC 794 (Shinde et al., 2020). We find a very good agreement for the wild-type and ΔgigC -mutant, but not for the Δgnd -mutant. The lower panel shows how the model predicts the allocation and consumption of each of the three strains.

Supplementary Table 1 | List of reactions added to iJC568 to form iSO595.

Supplementary Table 2 | List of reactions from iJC568 that are modified in iSO595.

Supplementary Table 3 | Parameter ranges used in the sampling of nutrient environments.

Supplementary Table 4 | List of blocked exchange reactions prior to sampling of nutrient environments.

Supplementary Table 5 | Parameter values used to run dynamic FBA in COMETS.

Supplementary Material 1 | Results from the BLAST-search used to identify 6PG-dehydratase (EC: 4.2.1.12) encoded by PMM0774 in *P. marinus* MED4.

Supplementary Material 2 | Memote snapshot report of iSO595.

- Arteaga, L., Pahlow, M., and Oschlies, A. (2016). Modeled Chl:C ratio and derived estimates of phytoplankton carbon biomass and its contribution to total particulate organic carbon in the global surface ocean. *Global Biogeochem. Cycles* 30, 1791–1810. doi: 10.1002/2016GB005458
- Becker, J. W., Berube, P. M., Follett, C. L., Waterbury, J. B., Chisholm, S. W., Delong, E. F., et al. (2014). Closely related phytoplankton species produce similar suites of dissolved organic matter. *Front. Microbiol.* 5:111. doi: 10.3389/fmicb.2014.00111
- Becker, J. W., and Hogle, S. L. (2019). Co-culture and biogeography of *Prochlorococcus* and SAR11. *ISME J.* 13, 1506–1519. doi: 10.1038/s41396-019-0365-4
- Benson, P. J., Purcell-Meyerink, D., Hocart, C. H., Truong, T. T., James, G. O., Rourke, L., et al. (2016). Factors altering pyruvate excretion in a glycogen storage mutant of the cyanobacterium, *synechococcus* PCC7942. *Front. Microbiol.* 7:475. doi: 10.3389/fmicb.2016.00475
- Bertilsson, S., Berglund, O., Pullin, M. J., and Chisholm, S. W. (2005). Release of dissolved organic matter by *Prochlorococcus*. *Vie et Milieu* 55, 225–231.
- Billler, S. J., Berube, P. M., Lindell, D., and Chisholm, S. W. (2015). *Prochlorococcus*: the structure and function of collective diversity. *Nat. Rev. Microbiol.* 13, 13–27. doi: 10.1038/nrmicro3378

- Billler, S. J., Coe, A., Roggensack, S. E., and Chisholm, W. (2018). Heterotroph interactions alter prochlorococcus transcriptome dynamics during extended periods of darkness. *mSystems* 3:e00040-18.
- Braakman, R. (2019). Evolution of cellular metabolism and the rise of a globally productive biosphere. *Free Radic. Biol. Med.* 140, 172–187. doi: 10.1016/j.freeradbiomed.2019.05.004
- Braakman, R., Follows, M. J., and Chisholm, S. W. (2017). Metabolic evolution and the self-organization of ecosystems. *Proc. Natl. Acad. Sci. U.S.A.* 114, E3091–E3100. doi: 10.1073/pnas.1619573114
- Bricaud, A., Claustre, H., Ras, J., and Oubelkheir, K. (2004). Natural variability of phytoplanktonic absorption in oceanic waters: influence of the size structure of algal populations. *J. Geophys. Res. Ocean* 109, 1–12. doi: 10.1029/2004JC002419
- Brodtrick, J. T., Rubin, B. E., Welkie, D. G., Du, N., Mih, N., Diamond, S., et al. (2016). Unique attributes of cyanobacterial metabolism revealed by improved genome-scale metabolic modeling and essential gene analysis. *Proc. Natl. Acad. Sci. U.S.A.* 113, E8344–E8353. doi: 10.1073/pnas.1613446113
- Burnap, R., Hagemann, M., and Kaplan, A. (2015). Regulation of CO₂ concentrating mechanism in cyanobacteria. *Life* 5, 348–371. doi: 10.3390/life5010348
- Campillo-Brocal, J. C., Lucas-Elió, P., and Sanchez-Amat, A. (2015). Distribution in different organisms of amino acid oxidases with fad or a quinone as cofactor and their role as antimicrobial proteins in marine bacteria. *Mar. Drugs* 13, 7403–7418. doi: 10.3390/md13127073
- Cano, M., Holland, S. C., Artier, J., Burnap, R. L., Ghirardi, M., Morgan, J. A., et al. (2018). Glycogen synthesis and metabolite overflow contribute to energy balancing in cyanobacteria. *Cell Rep.* 23, 667–672. doi: 10.1016/j.celrep.2018.03.083
- Carrieri, D., Paddock, T., Maness, P.-C., Seibert, M., and Yu, J. (2012). Photocatalytic conversion of carbon dioxide to organic acids by a recombinant cyanobacterium incapable of glycogen storage. *Energy Environ. Sci.* 5:9457. doi: 10.1039/c2ee23181f
- Casey, J. R., Mardinoglu, A., Nielsen, J., and Karl, D. M. (2016). Adaptive evolution of phosphorus metabolism in prochlorococcus. *mSystems* 1, 1–15. doi: 10.1128/mSystems.00065-16.Editor
- Chen, X., Schreiber, K., Appel, J., Makowka, A., Fähnrich, B., Roettger, M., et al. (2016). The Entner-Doudoroff pathway is an overlooked glycolytic route in cyanobacteria and plants. *Proc. Natl. Acad. Sci. U.S.A.* 113, 5441–5446. doi: 10.1073/pnas.1521916113
- Cirri, E., and Pohnert, G. (2019). Algae–bacteria interactions that balance the planktonic microbiome. *New Phytol.* 223, 100–106. doi: 10.1111/nph.15765
- Coles, V. J., Stukel, M. R., Brooks, M. T., Burd, A., Crump, B. C., Moran, M. A., et al. (2017). Ocean biogeochemistry modeled with emergent trait-based genomics. *Science* 358, 1149–1154. doi: 10.1126/science.aan5712
- Damrow, R., Maldener, L., and Zilliges, Y. (2016). The multiple functions of common microbial carbon polymers, glycogen and PHB, during stress responses in the non-diazotrophic cyanobacterium *Synechocystis* sp. PCC 6803. *Front. Microbiol.* 7:966. doi: 10.3389/fmicb.2016.00966
- Dauvillée, D., Kinderf, I. S., Li, Z., Kosar-Hashemi, B., Samuel, M. S., Rampling, L., et al. (2005). Role of the *Escherichia coli* glgX gene in glycogen metabolism. *J. Bacteriol.* 187, 1465–1473. doi: 10.1128/JB.187.4.1465-1473.2005
- Davey, M., Tarran, G. A., Mills, M. M., Ridame, C., Geider, R. J., and LaRoche, J. (2008). Nutrient limitation of picophytoplankton photosynthesis and growth in the tropical North Atlantic. *Limnol. Oceanogr.* 53, 1722–1733. doi: 10.4319/lo.2008.53.5.1722
- de Groot, D. H., Lischke, J., Muolo, R., Planqué, R., Bruggeman, F. J., and Teusink, B. (2020). The common message of constraint-based optimization approaches: overflow metabolism is caused by two growth-limiting constraints. *Cell. Mol. Life Sci.* 77, 441–453. doi: 10.1007/s00018-019-03380-2
- Deutsch, C., Sarmiento, J. L., Sigman, D. M., Gruber, N., and Dunne, J. P. (2007). Spatial coupling of nitrogen inputs and losses in the ocean. *Nature* 445, 163–167. doi: 10.1038/nature05392
- Diamond, S., Jun, D., Rubin, B. E., and Golden, S. S. (2015). The circadian oscillator in *Synechococcus elongatus* controls metabolite partitioning during diurnal growth. *Proc. Natl. Acad. Sci. U.S.A.* 112, E1916–E1925. doi: 10.1073/pnas.1504576112
- Díaz-Troya, S., López-Maury, L., Sánchez-Riego, A. M., Roldán, M., and Florencio, F. J. (2014). Redox regulation of glycogen biosynthesis in the cyanobacterium *Synechocystis* sp. PCC 6803: analysis of the AGP and glycogen synthases. *Mol. Plant* 7, 87–100. doi: 10.1093/mp/sst137
- Domínguez-Martín, M. A., López-Lozano, A., Clavería-Gimeno, R., Velázquez-Campoy, A., Seidel, G., Burkovski, A., et al. (2018). Differential NtcA responsiveness to 2-oxoglutarate underlies the diversity of C/N balance regulation in *Prochlorococcus*. *Front. Microbiol.* 8:2641. doi: 10.3389/fmicb.2017.02641
- Dubinsky, Z., and Berman-Frank, I. (2001). Uncoupling primary production from population growth in photosynthesizing organisms in aquatic ecosystems. *Aquatic Sci.* 63, 4–17. doi: 10.1007/PL00001343
- Dukovski, I., Bajić, D., Chacón, J. M., Quintin, M., Vila, J. C., Sulheim, S., et al. (2020). *COMETS: Computation of Microbial Ecosystems in Time and Space (COMETS): An Open Source Collaborative Platform for Modeling Ecosystems Metabolism*. Available online at: <http://arxiv.org/abs/2009.01734> (accessed December 17, 2020).
- Ebrahim, A., Lerman, J. A., Palsson, B. O., and Hyduke, D. R. (2013). COBRApy: constraints-based reconstruction and analysis for python. *BMC Syst. Biol.* 7:74. doi: 10.1186/1752-0509-7-74
- Edwards, J. S., Ramakrishna, R., and Palsson, B. O. (2002). Characterizing the metabolic phenotype: a phenotype phase plane analysis. *Biotechnol. Bioeng.* 77, 27–36. doi: 10.1002/bit.10047
- Elbourne, L. D. H., Tetu, S. G., Hassan, K. A., and Paulsen, I. T. (2017). TransportDB 2.0: a database for exploring membrane transporters in sequenced genomes from all domains of life. *Nucleic Acids Res.* 45, D320–D324. doi: 10.1093/nar/gkw1068
- Feist, A. M., Nagarajan, H., Rotaru, A. E., Tremblay, P. L., Zhang, T., Nevin, K. P., et al. (2014). Constraint-based modeling of carbon fixation and the energetics of electron transfer in geobacter metallireducens. *PLoS Comput. Biol.* 10:e1003575. doi: 10.1371/journal.pcbi.1003575
- Field, C. B., Behrenfeld, M. J., Randerson, J. T., and Falkowski, P. (1998). Primary production of the biosphere: integrating terrestrial and oceanic components. *Science* 281, 237–240. doi: 10.1126/science.281.5374.237
- Fischer, E., and Sauer, U. (2005). Large-scale in vivo flux analysis shows rigidity and suboptimal performance of *Bacillus subtilis* metabolism. *Nat. Genet.* 37, 636–640. doi: 10.1038/ng1555
- Flombaum, P., Gallegos, J. L., Gordillo, R. A., Rincón, J., Zabala, L. L., Jiao, N., et al. (2013). Present and future global distributions of the marine Cyanobacteria *Prochlorococcus* and *Synechococcus*. *Proc. Natl. Acad. Sci. U.S.A.* 110, 9824–9829. doi: 10.1073/pnas.1307701110
- Fogg, G. E., Nalewajko, C., and Watt, W. D. (1965). Extracellular products of phytoplankton photosynthesis. *Proc. R. Soc. London. Ser. B. Biol. Sci.* 162, 517–534. doi: 10.1098/rspb.1965.0054
- Follows, M. J., Dutkiewicz, S., Grant, S., and Chisholm, S. W. (2007). Emergent biogeography of microbial communities in a model ocean. *Science* 315, 1843–1846. doi: 10.1126/science.1138544
- Forchhammer, K., and Schwarz, R. (2019). Nitrogen chlorosis in unicellular cyanobacteria – a developmental program for surviving nitrogen deprivation. *Environ. Microbiol.* 21, 1173–1184. doi: 10.1111/1462-2920.14447
- Forchhammer, K., and Selim, K. A. (2019). Carbon/nitrogen homeostasis control in cyanobacteria. *FEMS Microbiol. Rev.* 44, 33–53. doi: 10.1093/femsre/fuz025
- Foster, S. Q., Al-haj, A., Church, M. J., van Dijken, G. L., Dutkiewicz, S., Fulweiler, R. W., et al. (2018). Ecological control of nitrite in the upper ocean. *Nat. Commun.* 9:1206. doi: 10.1038/s41467-018-03553-w
- Fu, J., and Xu, X. (2006). The functional divergence of two glgP homologues in *Synechocystis* sp. PCC 6803. *FEMS Microbiol. Lett.* 260, 201–209. doi: 10.1111/j.1574-6968.2006.00312.x
- Gilbert, J. D. J., and Fagan, W. F. (2011). Contrasting mechanisms of proteomic nitrogen thrift in *Prochlorococcus*. *Mol. Ecol.* 20, 92–104. doi: 10.1111/j.1365-294X.2010.04914.x
- Goelzer, A., Fromion, V., and Scorletti, G. (2011). Cell design in bacteria as a convex optimization problem. *Automatica* 47, 1210–1218. doi: 10.1016/j.automatica.2011.02.038
- Gomez, J. A., Höffner, K., and Barton, P. I. (2014). DFBAlab: a fast and reliable MATLAB code for dynamic flux balance analysis. *BMC Bioinform.* 15:409. doi: 10.1186/s12859-014-0409-8

- Grossowicz, M., Roth-Rosenberg, D., Aharonovich, D., Silverman, J., Follows, M. J., and Sher, D. (2017). *Prochlorococcus* in the lab and in silico: the importance of representing exudation. *Limnol. Oceanogr.* 62, 818–835. doi: 10.1002/lno.10463
- Gu, C., Kim, G. B., Kim, W. J., Kim, H. U., and Lee, S. Y. (2019). Current status and applications of genome-scale metabolic models. *Genome Biol.* 20, 1–18. doi: 10.1186/s13059-019-1730-3
- Gudmundsson, S., and Thiele, I. (2010). Computationally efficient flux variability analysis. *BMC Bioinform.* 11:489. doi: 10.1186/1471-2105-11-489
- Harcombe, W. R., Riehl, W. J., Dukovski, I., Granger, B. R., Betts, A., Lang, A. H., et al. (2014). Metabolic resource allocation in individual microbes determines ecosystem interactions and spatial dynamics. *Cell Rep.* 7, 1104–1115. doi: 10.1016/j.celrep.2014.03.070
- Haug, K., Salek, R. M., Conesa, P., Hastings, J., De Matos, P., Rijnbeek, M., et al. (2013). MetaLights—an open-access general-purpose repository for metabolomics studies and associated meta-data. *Nucleic Acids Res.* 41, 781–786. doi: 10.1093/nar/gks1004
- Heirendt, L., Arreckx, S., Pfau, T., Mendoza, S. N., Richelle, A., Heinken, A., et al. (2019). Creation and analysis of biochemical constraint-based models using the COBRA Toolbox v.3.0. *Nat. Protoc.* 14, 639–702. doi: 10.1038/s41596-018-0098-2
- Hennon, G. M., Morris, J. J., Haley, S. T., Zinser, E. R., Durrant, A. R., Entwistle, E., et al. (2018). The impact of elevated CO₂ on *Prochlorococcus* and microbial interactions with a ‘helper’ bacterium *Alteromonas*. *ISME J.* 12, 520–531. doi: 10.1038/ismej.2017.189
- Henry, C. S., Dejongh, M., Best, A. A., Frybarger, P. M., Linsay, B., and Stevens, R. L. (2010). High-throughput generation, optimization and analysis of genome-scale metabolic models. *Nat. Biotechnol.* 28, 977–982. doi: 10.1038/nbt.1672
- Holtzendorff, J., Partensky, F., Mella, D., Lennon, J.-F., Hess, W. R., and Garczarek, L. (2008). Genome streamlining results in loss of robustness of the circadian clock in the marine cyanobacterium *Prochlorococcus marinus* PCC 9511. *J. Biol. Rhythms* 23, 187–199. doi: 10.1177/0748730408316040
- Hopkinson, B. M., Young, J. N., Tansik, A. L., and Binder, B. J. (2014). The minimal CO₂-concentrating mechanism of *prochlorococcus* spp. MED4 is effective and efficient. *Plant Physiol.* 166, 2205–2217. doi: 10.1104/pp.114.247049
- Iglesias, A. A., Kafekuda, G., and Preiss, J. (1991). Regulatory and structural properties of the cyanobacterial ADP-glucose pyrophosphorylases. *Plant Physiol.* 97, 1187–1195. doi: 10.1104/pp.97.3.1187
- Johnson, Z. I., Zinser, E. R., Coe, A., McNulty, N. P., Woodward, E. M. S., and Chisholm, S. W. (2006). Niche partitioning among *Prochlorococcus* ecotypes along ocean-scale environmental gradients. *Science* 311, 1737–1740. doi: 10.1126/science.1118052
- Kanehisa, M., and Goto, S. (2000). KEGG: kyoto encyclopedia of genes and genomes. *Nucleic Acids Res.* 28, 27–30.
- Kashyap, A. K., and Singh, D. P. (1985). Ammonium transport in unicellular cyanobacterium *anacystis nidulans*. *J. Plant Physiol.* 121, 319–330. doi: 10.1016/S0176-1617(85)80025-0
- Kavvas, E. S., Seif, Y., Yurkovich, J. T., Norsigian, C., Poudel, S., Greenwald, W. W., et al. (2018). Updated and standardized genome-scale reconstruction of *Mycobacterium tuberculosis* H37Rv, iEK1011, simulates flux states indicative of physiological conditions. *BMC Syst. Biol.* 12:25. doi: 10.1186/s12918-018-0557-y
- Kettler, G. C., Martiny, A. C., Huang, K., Zucker, J., Coleman, M. L., Rodrigue, S., et al. (2007). Patterns and implications of gene gain and loss in the evolution of *prochlorococcus*. *PLoS Genet.* 3:e231. doi: 10.1371/journal.pgen.0030231
- Kim, J., Fabris, M., Baart, G., Kim, M. K., Goossens, A., Vyverman, W., et al. (2016). Flux balance analysis of primary metabolism in the diatom *Phaeodactylum tricornutum*. *Plant J.* 85, 161–176. doi: 10.1111/tpj.13081
- Knoop, H., Gründel, M., Zilliges, Y., Lehmann, R., Hoffmann, S., Lockau, W., et al. (2013). Flux balance analysis of cyanobacterial metabolism: the metabolic network of *Synechocystis* sp. PCC 6803. *PLoS Comput. Biol.* 9:e1003081. doi: 10.1371/journal.pcbi.1003081
- Krumhardt, K. M., Callnan, K., Roache-Johnson, K., Swett, T., Robinson, D., Reistetter, E. N., et al. (2013). Effects of phosphorus starvation versus limitation on the marine cyanobacterium *Prochlorococcus* MED4 I: uptake physiology. *Environ. Microbiol.* 15, 2114–2128. doi: 10.1111/1462-2920.12079
- Kujawinski, E. B. (2011). The impact of microbial metabolism on marine dissolved organic matter. *Ann. Rev. Mar. Sci.* 3, 567–599. doi: 10.1146/annurev-marine-120308-081003
- Laurenceau, R., Bliem, C., Osburne, M. S., Becker, J. W., Biller, S. J., Cubillos-Ruiz, A., et al. (2020). Toward a genetic system in the marine cyanobacterium *Prochlorococcus*. *Access Microbiol.* 2:e000107. doi: 10.1099/acmi.0.000107
- Lewis, N. E., Hixson, K. K., Conrad, T. M., Lerman, J. A., Charusanti, P., Polpitiya, A. D., et al. (2010). Omic data from evolved *E. coli* are consistent with computed optimal growth from genome-scale models. *Mol. Syst. Biol.* 6:390. doi: 10.1038/msb.2010.47
- Lieven, C., Beber, M. E., Olivier, B. G., Bergmann, F. T., Ataman, M., Babaei, P., et al. (2020). MEMOTE for standardized genome-scale metabolic model testing. *Nat. Biotechnol.* 38, 272–276. doi: 10.1038/s41587-020-0446-y
- Long, S. P., Humphries, S., and Falkowski, P. G. (1994). Photoinhibition of photosynthesis in nature. *Annu. Rev. Plant Physiol. Plant Mol. Biol.* 45, 633–662. doi: 10.1146/annurev.pp.45.060194.003221
- López-Sandoval, D. C., Rodríguez-Ramos, T., Cermeño, P., and Marañón, E. (2013). Exudation of organic carbon by marine phytoplankton: dependence on taxon and cell size. *Mar. Ecol. Prog. Ser.* 477, 53–60. doi: 10.3354/meps10174
- Luan, G., Zhang, S., Wang, M., and Lu, X. (2019). Progress and perspective on cyanobacterial glycogen metabolism engineering. *Biotechnol. Adv.* 37, 771–786. doi: 10.1016/j.biotechadv.2019.04.005
- Ma, L., Calfee, B. C., Morris, J. J., Johnson, Z. I., and Zinser, E. R. (2018). Degradation of hydrogen peroxide at the ocean's surface: the influence of the microbial community on the realized thermal niche of *Prochlorococcus*. *ISME J.* 12, 473–484. doi: 10.1038/ismej.2017.182
- Ma, X., Coleman, M. L., and Waldbauer, J. R. (2018). Distinct molecular signatures in dissolved organic matter produced by viral lysis of marine cyanobacteria. *Environ. Microbiol.* 20, 3001–3011. doi: 10.1111/1462-2920.14338
- Maarleveld, T. R., Khandelwal, R. A., Olivier, B. G., Teusink, B., and Bruggeman, F. J. (2013). Basic concepts and principles of stoichiometric modeling of metabolic networks. *Biotechnol. J.* 8, 997–1008. doi: 10.1002/biot.201200291
- Mague, T. H., Friberg, E., Hughes, D. J., and Morris, I. (1980). Extracellular release of carbon by marine phytoplankton; a physiological approach. *Limnol. Oceanogr.* 25, 262–279. doi: 10.4319/lo.1980.25.2.0262
- Mahadevan, R., Edwards, J. S., and Doyle, F. J. (2002). Dynamic Flux Balance Analysis of diauxic growth in *Escherichia coli*. *Biophys. J.* 83, 1331–1340. doi: 10.1016/S0006-3495(02)73903-9
- Marañón, E., Cermeño, P., López-Sandoval, D. C., Rodríguez-Ramos, T., Sobrino, C., Huete-Ortega, M., et al. (2013). Unimodal size scaling of phytoplankton growth and the size dependence of nutrient uptake and use. *Ecol. Lett.* 16, 371–379. doi: 10.1111/ele.12052
- Mary, I., Tu, C.-J., Grossman, A., and Vault, D. (2004). Effects of high light on transcripts of stress-associated genes for the cyanobacteria *Synechocystis* sp. PCC 6803 and *Prochlorococcus* MED4 and MIT9313. *Microbiology* 150, 1271–1281. doi: 10.1099/mic.0.27014-0
- McDonald, S. M., Plant, J. N., and Worden, A. Z. (2010). The mixed lineage nature of nitrogen transport and assimilation in marine eukaryotic phytoplankton: a case study of *Micromonas*. *Mol. Biol. Evol.* 27, 2268–2283. doi: 10.1093/molbev/msq113
- McInnes, L., Healy, J., and Astels, S. (2017). hdbscan: hierarchical density based clustering. *J. Open Source Softw.* 2:205. doi: 10.21105/joss.00205
- McKinney, W. (2010). *pandas: a Foundational Python Library for Data Analysis and Statistics | R (Programming Language) | Database Index*. Available online at: <https://www.scribd.com/document/71048089/pandas-a-Foundational-Python-Library-for-Data-Analysis-and-Statistics> (accessed April 29, 2020).
- Mella-Flores, D., Six, C., Ratin, M., Partensky, F., Boutte, C., Le Corguillé, G., et al. (2012). *Prochlorococcus* and *synechococcus* have evolved different adaptive mechanisms to cope with light and UV Stress. *Front. Microbiol.* 3:285. doi: 10.3389/fmicb.2012.00285
- Monk, J. M., Lloyd, C. J., Brunk, E., Mih, N., Sastry, A., King, Z., et al. (2017). iM1515, a knowledgebase that computes *Escherichia coli* traits. *Nat. Biotechnol.* 35, 904–908. doi: 10.1038/nbt.3956
- Monshupanee, T., and Incharoensakdi, A. (2014). Enhanced accumulation of glycogen, lipids and polyhydroxybutyrate under optimal nutrients and light intensities in the cyanobacterium *Synechocystis* sp. PCC 6803. *J. Appl. Microbiol.* 116, 830–838. doi: 10.1111/jam.12409
- Moore, C. M., Mills, M. M., Arrigo, K. R., Berman-Frank, I., Bopp, L., Boyd, P. W., et al. (2013). Processes and patterns of oceanic nutrient limitation. *Nat. Geosci.* 6, 701–710. doi: 10.1038/ngeo1765

- Moore, L. R., Coe, A., Zinser, E. R., Saito, M. A., Sullivan, M. B., Lindell, D., et al. (2007). Culturing the marine cyanobacterium *Prochlorococcus*. *Limnol. Oceanogr. Methods* 5, 353–362. doi: 10.4319/lom.2007.5.353
- Moore, L. R., Goericke, R., and Chisholm, S. W. (1995). Comparative physiology of *Synechococcus* and *Prochlorococcus*: influence of light and temperature on growth, pigments, fluorescence and absorptive properties. *Mar. Ecol. Prog. Ser.* 116, 259–275.
- Moradi, N., Liu, B., Iversen, M., Kuypers, M. M., Ploug, H., and Khalili, A. (2018). A new mathematical model to explore microbial processes and their constraints in phytoplankton colonies and sinking marine aggregates. *Sci. Adv.* 4, 1–10. doi: 10.1126/sciadv.aat1991
- Moran, M. A., and Durham, B. P. (2019). Sulfur metabolites in the pelagic ocean. *Nat. Rev. Microbiol.* 17, 665–678. doi: 10.1038/s41579-019-0250-1
- Moran, M. A., Kujawinski, E. B., Stubbins, A., Fatland, R., Aluwihare, L. L., Buchan, A., et al. (2016). Deciphering ocean carbon in a changing world. *Proc. Natl. Acad. Sci. U.S.A.* 113, 3143–3151. doi: 10.1073/pnas.1514645113
- Morel, A., and Bricaud, A. (1981). Theoretical results concerning light absorption in a discrete medium, and application to specific absorption of phytoplankton. *Deep Sea Res. Part A Oceanogr. Res. Pap.* 28, 1375–1393. doi: 10.1016/0198-0149(81)90039-X
- Morris, J. J., Lenski, R. E., and Zinser, E. R. (2012). The black queen hypothesis: evolution of dependencies through adaptive gene loss. *MBio* 3:e00036-12. doi: 10.1128/mBio.00036-12
- Muñoz-Marín, M. C., Gómez-Baena, G., López-Lozano, A., Moreno-Cabezuolo, J. A., Diez, J., and García-Fernández, J. M. (2020). Mixotrophy in marine picocyanobacteria: use of organic compounds by *Prochlorococcus* and *Synechococcus*. *ISME J.* 14, 1065–1073. doi: 10.1038/s41396-020-0603-9
- Nicholson, D. P., Stanley, R. H. R., and Doney, S. C. (2018). A Phytoplankton model for the allocation of gross photosynthetic energy including the trade-offs of diazotrophy. *J. Geophys. Res. Biogeosci.* 123, 1796–1816. doi: 10.1029/2017JG004263
- Nogales, J., Gudmundsson, S., Knight, E. M., Palsson, B. O., and Thiele, I. (2012). Detailing the optimality of photosynthesis in cyanobacteria through systems biology analysis. *Proc. Natl. Acad. Sci. U.S.A.* 109, 2678–2683. doi: 10.1073/pnas.1117907109
- Noreña-Caro, D., and Benton, M. G. (2018). Cyanobacteria as photoautotrophic biofactories of high-value chemicals. *J. CO₂ Util.* 28, 335–366. doi: 10.1016/j.jcou.2018.10.008
- O'Brien, E. J., Monk, J. M., and Palsson, B. O. (2015). Using genome-scale models to predict biological capabilities. *Cell* 161, 971–987. doi: 10.1016/j.cell.2015.05.019
- Orth, J. D., Thiele, I., and Palsson, B. O. (2010). What is flux balance analysis? *Nat. Biotechnol.* 28, 245–248. doi: 10.1038/nbt.1614
- Oschlies, A., Koeve, W., Landolfi, A., and Kähler, P. (2019). Loss of fixed nitrogen causes net oxygen gain in a warmer future ocean. *Nat. Commun.* 10, 1–7. doi: 10.1038/s41467-019-10813-w
- Pacheco, A. R., Moel, M., and Segrè, D. (2018). Costless metabolic secretions as drivers of interspecies interactions in microbial ecosystems. *Nat. Commun.* 10, 103.
- Park, J., and Choi, Y. (2017). Cofactor engineering in cyanobacteria to overcome imbalance between NADPH and NADH: a mini review. *Front. Chem. Sci. Eng.* 11:66–71. doi: 10.1007/s11705-016-1591-1
- Partensky, F., and Garczarek, L. (2010). *Prochlorococcus*: advantages and limits of minimalism. *Ann. Rev. Mar. Sci.* 2, 305–331. doi: 10.1146/annurev-marine-120308-081034
- Partensky, F., Hess, W. R., and Vault, D. (1999). *Prochlorococcus*, a marine photosynthetic prokaryote of global significance. *Microbiol. Mol. Biol. Rev.* 63, 106–127. doi: 10.1128/mmr.63.1.106-127.1999
- Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., et al. (2011). Scikit-learn: machine learning in python. *J. Mach. Learn. Res.* 12, 2825–2830.
- Pope, R. M., and Fry, E. S. (1997). Absorption spectrum (380–700 nm) of pure water. II. Integrating cavity measurements. *Appl. Optics* 33, 8710–8723. doi: 10.1364/AO.36.008710
- Prigent, S., Frioux, C., Dittami, S. M., Thiele, S., Larhlmi, A., Collet, G., et al. (2017). Meneco, a topology-based gap-filling tool applicable to degraded genome-wide metabolic networks. *PLoS Comput. Biol.* 13:e1005276. doi: 10.1371/journal.pcbi.1005276
- Reid, A. (2012). *Incorporating Microbial Processes into Climate Change Models*. A Report by the American Academy of Microbiology. Washington, DC.
- Reimers, A.-M., Knoop, H., Bockmayr, A., and Steuer, R. (2017). Cellular trade-offs and optimal resource allocation during cyanobacterial diurnal growth. *Proc. Natl. Acad. Sci. U.S.A.* 114:201617508. doi: 10.1073/pnas.1617508114
- Ribalet, F., Swallow, J., Clayton, S., Jiménez, V., Sudek, S., Lin, Y., et al. (2015). Light-driven synchrony of *Prochlorococcus* growth and mortality in the subtropical Pacific gyre. *Proc. Natl. Acad. Sci. U.S.A.* 112, 8008–8012. doi: 10.1073/pnas.1424279112
- Roth-rosenberg, D., Aharonovich, D., Omta, A.-W., Follows, M. J., and Sciences, P. (2019). Dynamic macromolecular composition and high exudation rates in *Prochlorococcus*. *bioRxiv* [Preprint]. doi: 10.1101/828897
- Rügen, M., Bockmayr, A., and Steuer, R. (2015). Elucidating temporal resource allocation and diurnal dynamics in phototrophic metabolism using conditional FBA. *Sci. Rep.* 5, 1–16. doi: 10.1038/srep15247
- Saito, M. A., McIlvin, M. R., Moran, D. M., Goepfert, T. J., DiTullio, G. R., Post, A. F., et al. (2014). Multiple nutrient stresses at intersecting Pacific Ocean biomes detected by protein biomarkers. *Science* 345, 1173–1177. doi: 10.1126/science.1256450
- Schuetz, R., Zamboni, N., Zampieri, M., Heinemann, M., and Sauer, U. (2012). Multidimensional optimality of microbial metabolism. *Science* 336, 601–604. doi: 10.1126/science.1216882
- Seaver, L. C., and Imlay, J. A. (2001). Hydrogen peroxide fluxes and compartmentalization inside growing *Escherichia coli*. *J. Bacteriol.* 183, 7182–7189. doi: 10.1128/JB.183.24.7182-7189.2001
- Segrè, D., Vitkup, D., and Church, G. M. (2002). Analysis of optimality in natural and perturbed metabolic networks. *Proc. Natl. Acad. Sci. U.S.A.* 99, 15112–15117. doi: 10.1073/pnas.232349399
- Sher, D., Thompson, J. W., Kashtan, N., Croal, L., and Chisholm, S. W. (2011). Response of *Prochlorococcus* ecotypes to co-culture with diverse marine bacteria. *ISME J.* 5, 1125–1132. doi: 10.1038/ismej.2011.1
- Shinde, S., Zhang, X., Singapuri, S. P., Kalra, I., Liu, X., Morgan-Kiss, R. M., et al. (2020). Glycogen metabolism supports photosynthesis start through the oxidative pentose phosphate pathway in cyanobacteria. *Plant Physiol.* 182, 507–517. doi: 10.1104/pp.19.01184
- Sosa, O. A., Casey, J. R., and Karl, D. M. (2019). Methylphosphonate oxidation in *Prochlorococcus* strain MIT9301 supports phosphate acquisition, formate excretion, and carbon assimilation into purines. *Appl. Environ. Microbiol.* 85:e00289-19. doi: 10.1128/AEM.00289-19
- Steglich, C., Behrenfeld, M., Koblizek, M., Claustre, H., Penno, S., Prasil, O., et al. (2001). Nitrogen deprivation strongly affects Photosystem II but not phycoerythrin level in the divinyl-chlorophyll b-containing cyanobacterium *Prochlorococcus marinus*. *Biochim. Biophys. Acta Bioenerg.* 1503, 341–349. doi: 10.1016/S0005-2728(00)00211-5
- Stramski, D., Bricaud, A., and Morel, A. (2001). Modeling the inherent optical properties of the ocean based on the detailed composition of the planktonic community. *Appl. Optics* 18, 2929–2945. doi: 10.1364/ao.40.002929
- Szul, M. J., Dearth, S. P., Campagna, S. R., and Zinser, E. R. (2019). Carbon fate and flux in *prochlorococcus* under nitrogen limitation. *mSystems* 4:e00254-18. doi: 10.1128/mSystems.00254-18
- Thiele, I., Fleming, R. M. T., Que, R., Bordbar, A., Diep, D., and Palsson, B. O. (2012). Multiscale modeling of metabolism and macromolecular synthesis in *E. coli* and its application to the evolution of codon usage. *PLoS One* 7:e45635. doi: 10.1371/journal.pone.0045635
- Thiele, I., Hyduke, D. R., Steeb, B., Fankam, G., Allen, D. K., Bazzani, S., et al. (2011). A community effort towards a knowledge-base and mathematical model of the human pathogen *Salmonella Typhimurium* LT2. *BMC Syst. Biol.* 5:8. doi: 10.1186/1752-0509-5-8
- Thiele, I., Vlassis, N., and Fleming, R. M. T. (2014). FASTGAPFILL: efficient gap filling in metabolic networks. *Bioinformatics* 30, 2529–2531. doi: 10.1093/bioinformatics/btu321
- Thornton, D. C. O. (2014). Dissolved organic matter (DOM) release by phytoplankton in the contemporary and future ocean. *Eur. J. Physiol.* 49, 20–46. doi: 10.1080/09670262.2013.875596
- van der Maaten, L., and Hinton, G. (2008). Visualizing data using t-SNE. *J. Mach. Learn. Res.* 9, 2579–2605.
- Van Mooy, B. A. S., Rocap, G., Fredricks, H. F., Evans, C. T., and Devol, A. H. (2006). Sulfolipids dramatically decrease phosphorus demand by

- picocyanobacteria in oligotrophic marine environments. *Proc. Natl. Acad. Sci. U.S.A.* 103, 8607–8612. doi: 10.1073/pnas.0600540103
- Varma, A., and Palsson, B. O. (1994). Metabolic flux balancing: basic concepts, scientific and practical use. *Bio/Technology* 12, 994–998. doi: 10.1038/nbt1094-994
- Vaulot, D., Marie, D., Olson, R. J., and Chisholm, S. W. (1995). Growth of *Prochlorococcus*, a photosynthetic prokaryote, in the equatorial Pacific Ocean. *Science* 268, 1480–1482. doi: 10.1126/science.268.5216.1480
- Virtanen, P., Gommers, R., Oliphant, T. E., Haberland, M., Reddy, T., Cournapeau, D., et al. (2020). SciPy 1.0: fundamental algorithms for scientific computing in Python. *Nat. Methods* 17, 261–272. doi: 10.1038/s41592-019-0686-2
- Waldbauer, J. R., Rodrigue, S., Coleman, M. L., and Chisholm, S. W. (2012). Transcriptome and proteome dynamics of a light-dark synchronized bacterial cell cycle. *PLoS One* 7:e43432. doi: 10.1371/journal.pone.0043432
- Ward, B. A., Collins, S., Dutkiewicz, S., Gibbs, S., Bown, P., Ridgwell, A., et al. (2019). Considering the role of adaptive evolution in models of the ocean and climate system. *J. Adv. Model. Earth Syst.* 11, 3343–3361. doi: 10.1029/2018MS001452
- Welkie, D. G., Rubin, B. E., Diamond, S., Hood, R. D., Savage, D. F., and Golden, S. S. (2019). A hard day's night: cyanobacteria in diel cycles. *Trends Microbiol.* 27, 231–242.
- Wintermute, E. H., Lieberman, T. D., and Silver, P. A. (2013). An objective function exploiting suboptimal solutions in metabolic networks. *BMC Syst. Biol.* 7:98. doi: 10.1186/1752-0509-7-98
- Xiong, W., Cano, M., Wang, B., Douchi, D., and Yu, J. (2017). The plasticity of cyanobacterial carbon metabolism. *Curr. Opin. Chem. Biol.* 41, 12–19. doi: 10.1016/j.cbpa.2017.09.004
- Yang, A. (2011). Modeling and evaluation of CO₂ supply and utilization in algal ponds. *Ind. Eng. Chem. Res.* 50, 11181–11192. doi: 10.1021/ie200723w
- Yoshikawa, K., Toya, Y., and Shimizu, H. (2017). Metabolic engineering of *Synechocystis* sp. PCC 6803 for enhanced ethanol production based on flux balance analysis. *Bioprocess Biosyst. Eng.* 40, 791–796. doi: 10.1007/s00449-017-1744-8
- Zavřel, T., Faizi, M., Loureiro, C., Poschmann, G., Stühler, K., Sinetova, M., et al. (2019). Quantitative insights into the cyanobacterial cell economy. *eLife* 8:e42508. doi: 10.7554/eLife.42508
- Zhang, C. C., Zhou, C. Z., Burnap, R. L., and Peng, L. (2018). Carbon/nitrogen metabolic balance: lessons from cyanobacteria. *Trends Plant Sci.* 23, 1116–1130. doi: 10.1016/j.tplants.2018.09.008
- Zinser, E. R., Lindell, D., Johnson, Z. I., Futschik, M. E., Steglich, C., Coleman, M. L., et al. (2009). Choreography of the transcriptome, photophysiology, and cell cycle of a minimal photoautotroph, *Prochlorococcus*. *PLoS One* 4:e5135. doi: 10.1371/journal.pone.0005135

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

Copyright © 2021 Ofaim, Sulheim, Almaas, Sher and Segrè. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

