RAFAEL AUGUSTO THEODORO PEREIRA DE SOUZA NAHAT

# Modelagem metabólica da produção de polihidroxialcanoatos com diferentes composições monoméricas por *Pseudomonas* sp.

Tese apresentada ao Programa de Pós-Graduação Interunidades em Biotecnologia da Universidade de São Paulo, Instituto Butantan e Instituto de Pesquisas Tecnológicas para obtenção do Título de Doutor em Biotecnologia

São Paulo
2019

RAFAEL AUGUSTO THEODORO PEREIRA DE SOUZA NAHAT

# Metabolic modelling of the production of polyhydroxyalkanoates with different monomeric compositions by *Pseudomonas* sp.

Thesis presented to the Interdepartamental Biotechnology Program of the Universidade de São Paulo, Instituto Butantan and Instituto de Pesquisas Tecnológicas for the degree of PhD in Biotechnology.

São Paulo
2019

RAFAEL AUGUSTO THEODORO PEREIRA DE SOUZA NAHAT

# Metabolic modelling of the production of polyhydroxyalkanoates with different monomeric compositions by *Pseudomonas* sp.

Thesis presented to the Interdepartamental Biotechnology Program of the Universidade de São Paulo, Instituto Butantan and Instituto de Pesquisas Tecnológicas for the degree of PhD in Biotechnology.

Area of concentration: Biotechnology

Supervisor: Dra Marilda Keico Taciro

Co-supervisor: Prof. Dr. José Gregório Cabrera Gomez

Original version

São Paulo
2019

RAFAEL AUGUSTO THEODORO PEREIRA DE SOUZA NAHAT

# Modelagem metabólica da produção de polihidroxialcanoatos com diferentes composições monoméricas por *Pseudomonas* sp.

Tese apresentada ao Programa de Pós-Graduação Interunidades em Biotecnologia da Universidade de São Paulo, Instituto Butantan e Instituto de Pesquisas Tecnológicas para obtenção do Título de Doutor em Biotecnologia

Área de concentração: Biotecnologia

Orientadora: Dra Marilda Keico Taciro

Coorientador: Prof. Dr. José Gregório Cabrera Gomez

Versão original

São Paulo
2019

CATALOGAÇÃO NA PUBLICAÇÃO (CIP)
Serviço de Biblioteca e informação Biomédica
do Instituto de Ciências Biomédicas da Universidade de São Paulo

Ficha Catalográfica elaborada pelo(a) autor(a)

Nahat, Rafael Augusto Theodoro Pereira de Souza
Modelagem metabólica da produção de polihidroxialcanoatos com diferentes composições monoméricas por Pseudomonas sp. / Rafael Augusto Theodoro Pereira de Souza Nahat; orientadora Marilda Keico Taciro; coorientador José Gregório Cabrera Gomez. -- São Paulo, 2019.
146 p.

Tese (Doutorado)) -- Universidade de São Paulo, Instituto de Ciências Biomédicas.

1. Biopolímeros. 2. Bioprocessos. 3. Otimização convexa. 4. Metabolismo celular. 5. Espaços vetoriais. I. Taciro, Marilda Keico, orientador. II. Gomez, José Gregório Cabrera, coorientador. III. Título.

UNIVERSIDADE DE SÃO PAULO
**Programa de Pós-Graduação Interunidades em Biotecnologia**
Universidade de São Paulo, Instituto Butantan, Instituto de Pesquisas Tecnológicas

Candidato(a): Rafael Augusto Theodoro Pereira de Souza Nahat

Titulo da Tese: Modelagem metabólica da produção de polihidroxialcanoatos com diferentes composições monoméricas por *Pseudomonas* sp.

Orientador: Dra. Marilda Keico Taciro .

A Comissão Julgadora dos trabalhos de Defesa da Tese de Doutorado, em sessão pública realizada a **........./......../..........**, considerou o(a) candidato(a):

**( ) Aprovado(a)** **( ) Reprovado(a)**

Examinador(a): Assinatura: ...............................................................................
Nome: .......................................................................................
Instituição: ...............................................................................

Examinador(a): Assinatura: ...............................................................................
Nome: .......................................................................................
Instituição: ...............................................................................

Examinador(a): Assinatura: ...............................................................................
Nome: .......................................................................................
Instituição: ...............................................................................

Examinador(a): Assinatura: ...............................................................................
Nome: .......................................................................................
Instituição: ...............................................................................

Examinador(a): Assinatura: ...............................................................................
Nome: .......................................................................................
Instituição: ...............................................................................

Presidente: Assinatura: ...............................................................................
Nome: .......................................................................................
Instituição: ...............................................................................

**UNIVERSIDADE DE SÃO PAULO**
**INSTITUTO DE CIÊNCIAS BIOMÉDICAS**

Cidade Universitária "Armando de Salles Oliveira"
Av. Prof. Lineu Prestes, 2415 – CEP. 05508-000 São Paulo, SP – Brasil
Telefone :(55) (11) 3091-7733 - telefax : (55) (11) 3091-8405
e-mail: cep@icb.usp.br

*Comissão de Ética em Pesquisa*

# CERTIFICADO DE ISENÇÃO

Certificamos que o Protocolo CEP-ICB Nº **638/14** referente ao projeto intitulado: "*Modelagem metabólica da produção do polihidroxialcanoato HB/HA MCL com diferentes composições monoméricas por Pseudomonas sp.*" sob a responsabilidade de **Rafael Augusto Theodoro Pereira de Souza Nahat,** foi analisado na presente data pela CEUA - Comissão de Ética no uso de animais e pela CEPSH- Comissão de Ética em Pesquisa com seres humanos, tendo sido deliberado que o referido projeto não utilizará animais que estejam sob a égide da lei 11.794 de 8 de outubro de 2008, nem envolverá procedimentos regulados pela Resolução CONEP nº466 de 2012.

São Paulo, 26 de fevereiro de 2014.

Prof. Dr. Wothan Tavares de Lima
Coordenador da CEUA - ICB/USP

Prof. Dr. Paolo M.A Zanotto
Coordenador da CEPsh - ICB/USP

# ACKNOWLEDGEMENTS

This work was funded by three different Brazilian institutions between 2014 and 2019, namely Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES), Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq) and Fundação de Amparo à Pesquisa do Estado de São Paulo (FAPESP). The periods and process numbers were:

- CAPES: 03/2014 to 06/2014 – no. 1364516 (Demanda Social, master's fellowship);
- CNPq: 03/2016 to 09/2016 – no. 140192/2016-6 (GM/GD-cotas, PhD fellowship);
- FAPESP: 10/2016 to 04/2019 – no. 2016/01756-0 (PhD fellowship);
- FAPESP: 06/2018 to 12/2018 – no. 2018/05456-6 (PhD internship in France);

First and foremost, I thank Prof. Ph.D. Aldo Tonso, who recommended me to the Laboratório de Bioprodutos in 2012 for the two last internship modules of my graduation course in Chemical Process Engineering at the Escola Politécnica da Universidade de São Paulo. Prof. Aldo was one of my internship supervisors, then the supervisor of my undergraduate thesis and also who encouraged me to pursue a Ph.D. degree straight after the graduation (skipping the master's degree) in order to continue the research in the field of polyhydroxyalkanoates bioprocessing.

Specifically about this thesis, I first thank Ph.D. Marilda Keico Taciro (the supervisor) and Profs. Ph.D. José Gregório Cabrera Gomes (the co-supervisor) and Luiziana Ferreira da Silva for the opportunity at the Laboratório de Bioprodutos. This work is a follow-up from some of their previous research. For example, Dr. Taciro's "Jovem Pesquisador" FAPESP project no. 2010/51692-1, which was the parent project of this work.

In second, I thank Dr. Juliana Cardinali Rezende, who carried-out laborious genome sequencing experiments for the bacterial strain *Pseudomonas* sp. LFM046, its functional annotation, the construction of mutant strains to test hypotheses on the metabolism of the LFM046 strain and transcriptome profiling experiments. She was also who started, in the Brazilian side, the cooperation between our research group and the French team European Research Team in Algorithms and Biology, formaL and Experimental (ERABLE), with whom the final consolidation of all the knowledge in this work was made possible.

During the six months of the internship in France with the ERABLE team (Université Claude Bernard Lyon 1 and Institut National de Recherche en Informatique et en Automatique), it was developed the *N*-phenotypic model and the first version of the software *N*-GlobalFit. This software is a proof-of-concept of an automated solution for one of the major bottlenecks of systems biology today, which is the refinement of genome-scale metabolic networks. These results are general and can help the future research of both groups.

In the ERABLE group, I thank the Ph.D. researchers Ricardo Luiz de Andrade Abrantes, Mariana Galvão Ferrarini, Marie-France Sagot, Alex di Genova, Arnaud Mary, Leandro Ishi Soares de Lima and Said Sadique Adi. Drs. Abrantes and Ferrarini were the closest direct collaborators in this internship in France, of which Dr. Sagot was the supervisor. The contact with this renowned group of specialists in mathematical modelling and computer science applied to bioinformatics added a whole new dimension of knowledge to this work, expanding its scope and impact. I cannot thank them enough for this opportunity.

Dr. Ferrarini was who started the collaboration between the Brazilian and French groups in the French side. Her and Dr. di Genova achieved the assembling of the full genome sequence of *Pseudomonas* sp. LFM046 for the very first time, using software previously developed by ERABLE. Prof. Ph.D. Mary gave advice on the mathematical formulation for the *N*-phenotypic model. Drs. de Lima and Adi gave general advice on mathematical formulation and software implementation.

Finally, I thank all the people not among the aforementioned colleagues but who also helped to carry-out the experiments whose results were directly used in this work, like the bioreactor cultures (some of which required sampling during the night), the analytic essays (e.g. chromatography), the enzyme activity essays and the construction of mutant and recombinant strains related to LFM046. Some of these people are: Drs. Karel Olavarria, Thatiane Teixeira Mendonça, Lucas Garbini Cespedes, MAs. Juan Camilo Roncallo Sarmiento, Edmar Ramos de Oliveira Filho, Henrique da Costa Oliveira and the laboratory technician Aelson Luiz dos Santos.

# RESUMO

NAHAT R. A. T. P. S. Modelagem metabólica da produção de polihidroxialcanoatos com diferentes composições monoméricas por *Pseudomonas* sp. [tese (Doutorado em Biotecnologia)] – Instituto de Ciências Biomédicas, Universidade de São Paulo, São Paulo, 2019

Uma grande base de dados gerados em quimiostato pela linhagem *Pseudomonas* sp. LFM046, produtora eficiente de polihidroxialcanoatos (PHA), estava disponível e foi utilizada para validar modelos metabólicos preditivos em escala genômica. Visando maximizar o teor de PHA com o controle da sua composição monomérica (e portanto das suas propriedades mecânicas), três etapas foram realizadas. Primeiro, um modelo *N*-fenotípico foi desenvolvido com base em 3 princípios biológicos e aplicado para selecionar os cultivos que melhor representavam todos os fenótipos possíveis da LFM046. Segundo, uma rede *draft* em escala genômica gerada automaticamente pelo *webservice* KBase a partir do genoma, montado pelo *software* MeDuSa, foi refinada usando dois métodos: o manual, que foi sistematizado e registrado para referência futura, e o *N*-GlobalFit, um novo método automático que foi validado como prova-de-conceito. As qualidades preditivas das redes metabólicas refinadas foram avaliadas usando o modelo *N*-fenotípico. Terceiro, simulações FBA (*Flux Balance Analysis*) específicas nessas redes confirmaram que a velocidade relativa da síntese *de novo* de ácidos graxos é afetada por condições de cultivo e determina o teor e a composição do PHA. Os métodos e *software* desenvolvidos neste trabalho, em colaboração com o grupo de pesquisa francês *European Research Team in Algorithms and Biology, formaL and Experimental* (ERABLE), foram generalizados a partir do estudo-de-caso da produção de PHA por *Pseudomonas* sp. LFM046 a fim de estabelecer uma *framework* de Pesquisa & Desenvolvimento (P&D) transferível para outras linhagens e bioprocessos.

Palavras-chave: Biopolímeros. Bioprocessos. Otimização convexa. Metabolismo celular. Espaços vetoriais.

# ABSTRACT

NAHAT R. A. T. P. S. Metabolic modelling of the production of polyhydroxyalkanoates with different monomeric compositions by *Pseudomonas* sp. [PhD thesis (Biotechnology)] – Instituto de Ciências Biomédicas, Universidade de São Paulo, São Paulo, 2019

A large database of chemostat cultures of the strain *Pseudomonas* sp. LFM046, an efficient polyhydroxyalkanoate (PHA) producer, was available and used to validate predictive genome-scale metabolic models. Aiming to maximize the PHA content while controlling its monomeric composition (and hence its mechanical properties), three steps were taken. First, an *N*-phenotypic model was developed based on three biological principles and then applied to select the cultivations which best represent all possible phenotypes of the strain LFM046. Second, a draft genome-scale reconstruction of LFM046 generated automatically by the webservice KBase from the genome, assembled by the software MeDuSa, was refined using two methods: the usual manual method, systematized and logged for future reference, and *N*-GlobalFit, a novel automated method which was validated as a proof-of-concept. The predictive qualities of the refined metabolic networks were assessed using the *N*-phenotypic model. Third, specific FBA (Flux Balance Analysis) simulations in these networks confirmed that the relative speed of the fatty-acids *de novo* synthesis is affected by culture conditions and determines the PHA content and composition. The methods and software developed in this work, in collaboration with the French research group European Research Team in Algorithms and Biology, formaL and Experimental (ERABLE), were generalized from the case-study of PHA production by *Pseudomonas* sp. LFM046 in order to stablish an R&D framework transferable to other strains and bioprocesses.

Keywords: Polyhydroxyalkanoates. Bioprocess optimization. Flux Balance Analysis. Genome-scale metabolic model. Predictive phenotypic model.

# INDEX OF FIGURES

# INDEX OF TABLES

# INDEX OF ABBREVIATIONS AND ACRONYMS

3HB – 3-hydroxybutyrate

ADP – adenosine-diphosphate

AEDC – average euclidean distance per coordinate

AMRE – average magnitude of the relative error

ARE – average relative error

ATP – adenosine-triphosphate

BIOTOT – total biomass (cells + intracellular PHA)

BLAST – basic local alignment search tool

$C_{10}$ – 10-carbon fatty-acid intermediary metabolite

$C_{12}$ – 12-carbon fatty-acid intermediary metabolite

$C_{12\Delta5}$ – 12-carbon fatty-acid intermediary metabolite, unsaturated in carbon 5

$C_4$ – 4-carbon fatty-acid intermediary metabolite

$C_6$ – 6-carbon fatty-acid intermediary metabolite

$C_8$ – 8-carbon fatty-acid intermediary metabolite

CPLEX – a software solver manufactured by IBM

DDR3 – double data rate type 3 (a type of RAM memory)

dFBA – dynamic flux balance analysis

DNA – deoxyribonucleic acid

DRUM – dynamic reduction of unbalanced metabolism

*E. coli* – the bacterial species *Escherichia coli*

*e*-value – error tolerance for the BLAST software

EC number – Enzyme Commission number (identification standard)

EDC – energy-dissipating cycle

EDEMP – Entner-Doudoroff and Embden-Meyerhoff-Parnas supercycle (metabolic pathway)

EFMA – elementary flux mode analysis

EFV – elementary flux vector

EGC – energy-generating cycle

EPirt – extended Pirt (maintenance energy model)

ERABLE – european research team in algorithms and biology, formal and experimental

EU – European Union

F.A.M.E. – flux analysis and modelling environment

FASTA – fast-all file format (“all” because it can be used for proteins or nucleotides)

FBA – flux balance analysis

FCA – flux coupling analysis

fru – fructose

FVA – flux variability analysis

GHz – gigahertz (unit of measurement of frequency)

glu – glucose

GPR – gene-protein-reaction association

HA – 3-hydroxyalkanoate

$HA_{MCL}$ – medium-chain-length 3-hydroxyalkanoate (6 to 14 carbons)

$HA_{SCL}$ – short-chain-length 3-hydroxyalkanoate (2 to 5 carbons)

HCM – hybrid cybernetic modelling

IBM – international business machines, the manufacturer of the software CPLEX

ID – identification number

iPAE1146 – a metabolic model of the bacterial strain *Pseudomonas aeruginosa* PAO1

IPT046 – deprecated name of the bacterial strain *Pseudomonas* sp. LFM046

KBase – United States department of energy systems biology knowledge base

KT2440 – a bacterial strain of the species *Pseudomonas putida*

LFM046 – a bacterial strain of the genus *Pseudomonas*

LP – linear programming or linear problem (optimization)

Lumped-HCM – lumped hybrid cybernetic modelling

MBMA – macroscopic biorreaction models analysis

MCL – medium-chain-length (for hydroxyalkanoates)

MeDuSa – Multi-Draft based Scaffolder (software for assembling genome sequences)

METANETX – meta-database of reaction and metabolite identification numbers

MFA – metabolic flux analysis

MILP – mixed-integer linear programming or mixed-integer linear problem (optimization)

MO – multi-objective

MO-FBA – multi-objective flux balance analysis

MO-FVA – multi-objective flux variability analysis

ModelSEED – interface for building metabolic models of the SEED environment

MPE – maximum predictive error

$NAD^+$ – nicotinamide adenine dinucleotide (oxidised form)

NADH – nicotinamide adenine dinucleotide (reduced form)

$NADP^+$ – nicotinamide adenine dinucleotide phosphate (oxidised form)

NADPH – nicotinamide adenine dinucleotide phosphate (reduced form)

OBS – observation

OR – logical operator “or” (for true/false statements)

P(3H5PV) – poly-3-hydroxy-5-phenylvaleric acid

P(3HB:$HA_{MCL}$) – copolymer of 3-hydroxybutyrate and medium-chain-length hydroxyalkanoates

P(3HB:HHx) – copolymer of 3-hydroxybutyrate and 3-hydroxyhexanoate

P(3HB:3HV:3HHx) – copolymer of 3-hydroxybutyrate, 3-hydroxyvalerate and 3-hydroxyhexanoate

P(3HB:3HP) – copolymer of 3-hydroxybutyrate and 3-hydroxypentanoate

P(3HB:3HV:4HB) – copolymer of 3-hydroxybutyrate, 3-hydroxyvalerate and 4-hydroxybutyrate

P(3HB:3HV) – copolymer of 3-hydroxybutyrate and 3-hydroxyvalerate

P(3HB:4HB) – copolymer of 3- and 4-hydroxybutyrate

P(3HHp) – poly-3-hydroxyheptanoate

P(3HHx:3HO:3HD:3HDD) – copolymer of 3-hydroxyhexanoate, 3-hydroxyoctanoate, 3-hydroxydecanoate and 3-hydroxydodecanoate

P(3HHx:3HO:3HD) – copolymer of 3-hydroxyhexanoate, 3-hydroxyoctanoate, 3-hydroxydecanoate

P(3HHx) – poly-3-hydroxyhexanoate

P(3HN) – poly-3-hydroxynonanoate

P(3HO) – poly-3-hydroxyoctanoate

P(3HV) – poly-3-hydroxyvalerate

P(4HB) – poly-4-hydroxybutyrate

P($HA_{SCL}$:$HA_{MCL}$) – copolymer of short-chain-length and medium-chain-length hydroxyalkanoates

PAO1 – a bacterial strain of the species *Pseudomonas aeruginosa*

PCA – principal component analysis

pFBA – parsimonious enzyme usage flux balance analysis

PHA – polyhydroxyalkanoate (or polyhydroxyalkanoates)

$PHA_{MCL}$ – medium-chain-lenght polyhydroxyalkanoates

$PHA_{SCL}$ – short-chain-lenght polyhydroxyalkanoates

PHB – polyhydroxybutyrate

PLA – polylactate

PpuQY1140 – a metabolic model of the bacterial strain *Pseudomonas putida* KT2440

PRIAM – enzyme-specific profiles for genome annotation (a database and software)

R&D – research and development

RAM – random access memory (for computers)

RNA – ribonucleic acid

SCL – short-chain-length (for hydroxyalkanoates)

Tol – tolerance

uFBA – unsteady flux balance analysis

USA – United States of America

# INDEX OF SYMBOLS AND CHEMICAL SPECIES

€ – euro (currency of the European Union)

$\vec{0}$ – column-vector of zeroes (null vector)

$a_H$ – constant endogenous metabolism rate of the Herbert model (Wang & Post, 2012)

c – a weight, error or tolerance; specific supply rate

C – Carbon (chemical element)

$C$ – collection of curves

$C_i$ – the i-th curve

$\mathrm{C_M}$ – fatty-acids intermediary metabolite with $M$ carbons

D – dilution rate

$F$ – liquid total flow rate; or fructose concentration

$f_S$ – a sigmoid manifold of the limitation regime S-dim(S)-lim in N-Liebscher space

$G$ – glucose concentration

H – Hydrogen (chemical element)

$H_p$, $H_{pp}$ and $H_{pn}$ – (hyper)plane and its defining point and normal vector, respectively

$k$ – a generic kinetic parameter

$K, p, r$ – shape parameters for the $N$-phenotypic model

$K_S$ – Monod constant parameter (Monod, 1949)

$m$ – constant maintenance energy term of the Pirt model

$\vec{m}'(t)$ – column-vector of all derivatives (with respect to time) of masses of metabolites, functions of time

$\vec{m}(t)$ – column-vector of all masses of metabolites, functions of time

$m_I$ and $m'$ – parameters of the extended Pirt (EPirt) model (Pirt, 1982)

$n$ – a natural number

N – Nitrogen (chemical element)

$N$ – number of independent non-interchangeable growth-limiting factors

$O_2$, $NH_4^+$, $PO_4^{3-}$, $CO_2$ – chemical species denoted by their chemical formulas

$P$ – a point

P – Phosphorus (chemical element)

$P\Sigma$ – sigmoid axis of the $N$-phenotypic model

$q$ – specific uptake or production rate

$Q$ and $Q'$ – two lines

$\vec{r}$ – column-vector of reaction rates, functions of time

$R_i$ – the *i*-th region

$S$ – non-normalized concentration of a generic substrate

S – Sulfur (chemical element); a set of substrates; or a generic substrate

Sr – set of reference substrates (can be a single substrate)

$t$ – time

$T_{g,k}$ – the threshold of growth of the *k*-th phenotypic point

$T_{h,k}$ – the threshold of non-growth of the *k*-th phenotypic point

US$ – United States dollars (currency of the United States of America)

$\vec{v}$ – column-vector of metabolic fluxes

V – volume

$v_j$ – the metabolic flux of the *j*-th reaction

$w$ – a weight parameter

$x$ – a generic variable

X – biomass; mass of biomass; or concentration of biomass

$y$ – molar or mass fraction

$Y_{A/B}$ – yield from B to A; ratio between a variable of A and the same variable of B

$Y_G$ – maximum biomass yield from a single limiting substrate S

$\vec{Z}$ – vector of normalized uptake rates

$Z_i$ – normalized uptake rate of the *i*-th substrate (Si)

$\alpha$ – growing fraction of the population (1- α is the dormant fraction) (Pirt, 1987)

$\alpha_1$ – position in the sigmoid axis of the *N*-phenotypic model

$\alpha_2$ – angular position around the sigmoid axis of the *N*-phenotypic model

$\gamma$ – a random number

$\delta$ – a weight, error, tolerance or uncertainty

$\epsilon$ – an error or a tolerance

$\lambda$ – stoichiometric matrix

$\lambda_i$ – *i*-th stoichiometric coefficient

$\mu$ – specific growth rate

$\mu_{max}$ – maximum specific growth rate

$\mu_{max,H}$ – maximum specific growth rate of the Herbert model (Wang & Post, 2012)

$\mu_{max,P}$ – maximum specific growth rate of the Pirt model (Wang & Post, 2012)

$\mu_{min}$ – minimum specific growth rate

OBS: this index is not exhaustive. Units of physical quantities (g, h, mL, etc) and local variables used merely for breaking large equations in smaller pieces were omitted.

# SUMMARY

# 1 INTRODUCTION

Natural resources are finite and even those constantly recycled in the biosphere are not abundant in a useful form everywhere in the world (e.g. process water). The larger the human population, the more important sustainable development becomes. In the industry, this means efficiency improvement: less waste, less residues, more recycling and better supply chains.

A bioprocess is a production process powered by a biological system such as a microbial community or an enzyme. Replacing a traditional process by a bioprocess can eliminate toxic residues such as organic solvents, decrease energy consumption due to milder conditions (e.g. temperature), replace fossil by renewable feedstock (e.g. petroleum by sugar cane) and increase reaction selectivity/specificity (e.g. an enzyme acting on the individual molecular level vs. a catalyst acting on a bulk mass level).

Therefore, developing competitive bioprocesses is one of the strategies to pursue sustainable development. An example is to replace traditional mining processes by a bioleaching alternative process (Bobadilla-Fazzini et al, 2017) which has been developed and optimized with the help of predictive models of the complex biological systems employed (Latorre et al, 2016). Such models are an ubiquitous need in biotechnology research and development (biotech R&D).

The word “bioeconomy” is commonly used to describe the shift from traditional processes to bio-based circular (closed-loop) alternatives on a global level (Bugge et al, 2016). It is already a large market, estimated at € 2 trillion and 17 million jobs in the EU (Ronzon et al, 2015), and US$ 393 billion and 4.2 million jobs in the USA (Golden et al, 2018). In Brazil, the sucroenergetic sector alone produces a gross yearly turnover of US$ 100 billion and 2.4 million jobs (União da Indústria de Cana-de-Açúcar & Centro Nacional das Indústrias do Setor Sucroenergético e Biocombustíveis, 2016; Castro et al, 2018).

In order to reduce the risks of that shift, the industrial units of the bioeconomy are being designed to be flexible. These “biorefineries” integrate several complementary bioprocesses to transform otherwise residues into feedstock and simplify supply chains (Santos et al, 2018; Hassan et al, 2019). An example of the need for this flexibility is what happened with biofuels in the last decade: once thought to be the definitive answer to sustainable development (Ernsting & Smolke,  2018), they are now known to compete with food production (Wiggins, 2010) and to have less potential of generating jobs and adding value than bioplastics (Organization for Economic Co-operation and Development, 2013), which in turn are now also raising the same questions (Bolger, 2018; Ißbrücker, 2018).

This change from biofuels to bioplastics is also a recent example of a major change in the demand for biotech R&D. There is strong evidence that this kind of change is not isolated and, moreover, is likely to occur at an increasingly faster rate in all sectors (Kurzweil, 2004). Thus, the time of response of a biotech R&D institution is increasingly more crucial.

One way to reduce this time is to minimize the risk of facing a R&D demand for which there are no promising starting points readily available. The same idea of flexibility from biorefineries can be applied by having a set of biological systems with complementary potential applications. But how to define this complementarity and evaluate it before knowing anything about the application itself and before having the results of many experiments?

An answer is to make use of predictive mathematical models of each biological system candidate to predict their possible phenotypes and classify these phenotypes in groups of promising applications. For the R&D of bioprocesses powered by alive microbial populations (a field within biotech R&D), these models are named “metabolic models”. The simplest kind of metabolic model is a “(metabolic) network”, which is a set of algebraic equations representing a set of biochemical reactions in steady-state. The rate of a reaction is deemed “(metabolic) flux” and the set of all fluxes of a network is a “flux distribution”.

The most interconnected reaction in a metabolic network is the “biomass reaction”, which produces biomass. This reaction is an artificial reaction in the sense that it does not actually exist but it is a way to constrain the biomass chemical composition of the possible flux distributions predicted by the model. A “genome-scale” metabolic network is one in which each non-artificial reaction is explicitly associated to a minimal set of genes, and this is deemed “Gene-Protein-Reaction association”, or GPR. These genome-scale networks usually have thousands of reactions. Flux Balance Analysis (FBA) is the standard simulation technique to predict phenotypes from genome-scale metabolic networks (Terzer et al, 2009).

In this context, an example of complementary potential applications are two strains such that for the same unit flux of production of Acetyl-CoA, one produces a high flux of reducing power equivalents (e.g. NADPH) whereas the other produces a low flux: the first is more likely to be an efficient at producing reduced biomolecules whereas the second is likely efficient at the producing oxidised biomolecules. Naturally, the reliability of this evaluation depends on the metabolic network used to predict the flux distributions.

A genome-scale network tends to be represent better a metabolism than a simplified network. For example, if the true metabolism has 1000 reactions and its model has only 50, some of the equations are combinations of the true equations, what is effectively to assume that part of the flux distribution is constant. Therefore, generating genome-scale metabolic networks in a reproducible and automated way is key for bioprocess R&D.

One bioprocess of industrial interest is the production of polyhydroxyalkanoates (PHA) by a *Pseudomonas* strain. PHA is a family of polyesters which can be synthesized from renewable sources and accumulated in intracellular granules by some microbial strains (Anderson & Dawes, 1990).

There are 150 possible different PHA monomers described in literature (Rhem, 2003). Like proteins, the PHA structure can also be classified in four levels: (i) primary: linear order of the distinct monomers within a PHA molecule; (ii) secondary: crystalline and amorphous repeating domains within a molecule due to its folding; (iii) tertiary: overall 3D shape of a PHA molecule; (iv) quaternary: overall 3D shape of multiple PHA molecules assembled together via intermolecular interactions.

PHA is biocompatible and this property depends mostly on the primary structure: the chemical nature of the monomers, including eventual ligands or molecules from other families of compounds, molecular weight and surface properties of PHA films (Shrivastav et al, 2013). This material is also biodegradable and that is dependent on the primary structure, the macroscopic shape (e.g. a bottle, a pellet or a film) and the species of the microbial community available at the place of disposal (Numata et al, 2009).

A great advantage of PHA over other materials deemed biodegradable is that PHA does not require specific composting/biodegrading industrial facilities: PHA is fully biodegradable in regular soil (Shrivastav et al, 2011; Altaee et al, 2016) and also in the sea (Greene, 2012; Sashiwa et al, 2018), despite the limitations of the current standards of marine biodegradability essays (Harrison et al, 2018).

The mechanical properties such as tensile strength and glass transition temperature depend mostly on the secondary structure: short-chain-length PHA monomers ($HA_{SCL}$) have 2 to 5 carbon atoms in their main chain and form crystalline lattices which make the material thermorigid and brittle, whereas medium-chain-length monomers ($HA_{MCL}$) have 6 to 14 carbons and form amorphous structures which have the elastomeric behaviour of flexible rubbers (Sudesh et al, 2000).

Finally, the tertiary and quaternary structures are the most important for the overall mechanical and physicochemical properties of PHA blends and copolymers. For example, for a copolymer $P(HA_{SCL}:HA_{MCL})$ with a given proportion between SCL and MCL monomers in the same molecule, a random monomeric distribution is unlikely to form as many crystalline domains than a block distribution, where the SCL monomers end up close together in tridimensional space after the folding and assembly of the molecules.

Despite the variety of isolated HA monomers described in literature, only a small number of homopolymers and copolymers of PHA have been reported (some of which are produced by recombinant strains). In most cases, $PHA_{SCL}$ are produced, such as PHB, P(3HB:3HV), P(3HB:4HB), P(3HB:3HP) and P(3HB:3HV:4HB). Some contain $HA_{MCL}$, e.g. P(3HB:3HV:3HHx), P(3HHx:3HO:3HD) and P(3HHx:3HO:3HD:3HDD). The less frequent group are the PHA homopolymers without 3HB, e.g. P(4HB), P(3HV), P(3H5PV), P(3HHx), P(3HHp), P(3HO) and P(3HN) (Singh et al, 2015).

Thus, combining different monomers via genetic modification of microbial strains and/or controlling cultivation conditions is a strategy to obtain PHA copolymers of convenient tertiary and quaternary structures. Then, to control this structure is one way to make a PHA bioprocess flexible enough for the increasingly faster pace of change in society’s demands.

The current main challenge for the market development of PHA as an industrial material is its relatively high production cost (Możejko-Ciesielska & Kiewisz, 2016). Besides efforts to reduce this cost, the control of the monomeric composition allows the production of tailored materials for high-value applications (e.g. healthcare) that can absorb it.

To achieve this, an accurate predictive genome-scale metabolic model of an efficient PHA-producing strain can be used to consolidate all the knowledge and data available about this strain, and then to carry-out *in silico* experiments to find possible optimization strategies for a bioprocess using this strain to produce PHA with control of the monomeric composition.

# 2 OBJECTIVES

The general objective of this work was to find optimization strategies for the production of polyhydroxyalkanoates (PHA) by *Pseudomonas* sp. LFM046 and/or recombinant strains based on it, envisioning the maximization of the intracellular PHA content with control of the monomeric composition. To achieve that, the following specific objectives were defined:

- build a genome-scale metabolic model of *Pseudomonas* sp. LFM046 which consolidates all the knowledge available about this strain, conciliating all conflicting information which is possible;
- validate this genome-scale metabolic model with all the available bioreactor experiments of *Pseudomonas* sp. LFM046;
- determine a range of maximum theoretical PHA content in function of the monomeric composition and pinpoint strategies to maximize the PHA content and at the same time manipulate the monomeric composition, by means of *in silico* experiments using the genome-scale metabolic model. The strategies may include genetic modification and also control of the bioprocess conditions (e.g. composition of the culture medium).

A secondary general objective was added with the collaboration with the French research group European Research team in Algorithms and Biology, formaL and Experimental (ERABLE): to generalize the methodological pipeline used in the aforementioned specific objectives, in order to stablish it as a routine R&D process in the Laboratório de Bioprodutos. The aim is to apply this pipeline for other strains and case-studies after the production of PHA by *Pseudomonas* sp. LFM046 and recombinants of it.

# 3 LITERATURE REVIEW

Given the objectives of (i) obtaining an accurate predictive genome-scale metabolic model of an efficient PHA-producing strain; and (ii) to carry-out *in silico* experiments with this model targeted at finding possible optimization strategies for a bioprocess using such strain to produce PHA with the control of the monomeric composition; the following is discussed in this specific literature review:

I. on the objectives themselves:

- why PHA instead of any other material,
- why a bioprocess powered by alive cells instead of possibly simpler process using catalysts or isolated enzymes,
- why the bacterial strain *Pseudomonas* sp. LFM046 instead of any other.

II. on the experimental data:

- genomic data and state-of-the-art genome assembly approaches,
- phenotypic data and methods of data reconciliation/outlier detection,
- empirical and theoretical relationships within the phenotypic data,
- kinetic and stoichiometric aspects of phenotypic data.

III. on the genome-scale metabolic models:

- stoichiometric and kinetic metabolic models,
- state-of-the-art methods of construction, refining and validation,
- definitions of maintenance energy and methods of modelling it.

IV. on the *in silico* experiments:

- state-of-the-art simulation techniques for metabolic models,
- the case for an approach based on Flux Balance Analysis (FBA).

## 3.1 On the objectives themselves

As mentioned in section 1, polyhydroxyalkanoates (PHA) have better biodegradability than other materials marketed as such, like polylactide (PLA), specially in the marine environment. Also, with 150 different monomers there are 11175 copolymers with 2 equal-sized blocks, which are just a very small fraction of all the possible combinations with significant differences in their mechanical properties. Thus it is a very versatile material.

There are applications for PHA with low molecular weights, but breaking a large molecule into smaller ones is easier than the opposite, as it can even happen undesirably during processing (Bugnicourt et al, 2014). Moreover, the mechanical strength of PHA is positively correlated to its molecular weight, and achieving high weights *in vitro* with high yields still depends on expensive purified reactants (Tsuge, 2016). Thus, a microbial strain is still the simplest kind of biological system suitable to develop a PHA bioprocess.

The wildtype bacterial strain *Pseudomonas* sp. LFM046 accumulates $PHA_{MCL}$ from glucose and fructose with yields among the highest ever reported (Poblete-Castro et al, 2014). It has been isolated from the rhizosphere of sugar cane (Gomez, 1994), found to accumulate $PHA_{MCL}$ when grown under nitrogen and/or phosphorous limitation (Gomez, 2000; Sánchez et al, 2003), tested in high-density fed-batch conditions (Diniz et al, 2004), characterized in over 3200 h of experiments in chemostat that were consolidated in a simplified metabolic model (Taciro, 2008) and its PHA production was compared to other similar wildtype strains (Silva-Queiroz et al, 2009).

Then, this strain was selected as a platform for the construction of recombinant strains using plasmids and antibiotics targeting the control of the PHA monomeric composition (Gomes, 2009), had its internal metabolism studied in depth with labelled carbon ($^{13}C$) experiments (Riascos et al, 2013), had its full genome sequenced (Cardinali-Rezende et al, 2015), originated a recombinant strain which produced a non-naturally-occurring $P(3HB\text{:}HA_{MCL})$ copolymer (Cespedes et al, 2018) and finally was successfully modified directly in its genome, eliminating the stability problems of plasmids and the need for antibiotics in the culture medium (Oliveira-Filho et al, 2018). Therefore, not only LFM046 is a flexible and efficient PHA producing strain as there is an extensive knowledge base about it.

## 3.2 On the experimental data

The genomic data of *Pseudomonas* sp. LFM046 is a set of 3,700,436 forward and reverse sequence reads containing the information of 5,970,318 base pairs and 5440 coding sequences which were annotated. (Cardinali-Rezende et al, 2015). This large dataset was generated using second-generation sequencing, specifically the MiSeq platform.

This kind of sequencing technology is the most used today, after it dramatically reduced the sequencing costs in 2007 (National Human Genome Research Institute, 2018). The idea is to copy the genome many times, break the copies in random positions such that fragments (reads) are small enough, sequence the fragments and finally assemble the whole sequence by evaluating the probabilities of each set of identical sequences to correspond to the same overlapping part of the genome (Besser et al, 2018).

The main improvement of the third generation over the second is to increase the length of the reads. This solves several shortcomings of the second generation such as the difficulty to correctly determine the position of repeated sequences in the genome (multiple gene copies), the inconclusiveness of regions with many repetitions of short sequences (short tandem repeats), the inaccuracy of annotations when sequencing RNA and the limited detection of epigenetic modifications (Van Dijk et al, 2018).

Therefore, from the second-generation sequencing onwards, the final assembly step is done in increasingly more powerful software (Muir et al, 2016). There are several algorithms, all bounded by theoretical limits like a minimum length of the reads (Bresler et al, 2013). Also, besides comparing the assembled genome to sequences of other organisms, these sequences can be used to disambiguate and fill gaps during the assembly (Bosi et al, 2015). As the field evolves to the use of long-reads (third generation sequencing), the computational problem of the assembly becomes much more complex, but it is possible to greatly simplify it by using a hybrid method that uses both long- and short-reads (Di Genova et al, 2018). That could prevent short-read sequencing from becoming obsolete with the third generation.

The assembled genome is then input in a software that generates an automatic draft reconstruction, which is a draft genome-scale metabolic network. This draft reconstruction is a mathematical problem completely independent of that of the assembly.

### 3.2.1 The phenotypic data

According to section 3.2, the genomic experimental data is used to generate an assembled genome sequence which in turn is used to generate a draft genome-scale metabolic network. But that is only one dimension of the whole problem of building the final network. The other dimension is the set of all possible phenotypes of the organism, which cannot be directly inferred from the genome due to two reasons: (i) the relationship between the set of genes and the set of possible biochemical reactions is not a simple one-to-one mapping; and (ii) the actual phenotype exhibited by an organism is a function also of the environmental conditions it is in, e.g. temperature, concentrations, phase in its life cycle, etc.

Therefore, in order to assess, validate or refine a metabolic model, a phenotypic dataset is required. For example, if *in silico* simulations of the draft model predict growth on glucose but not on fructose as sole carbon sources whereas *in vivo* the organism grows on both, then the model lacks reactions related to fructose metabolization. The simplest type of phenotypic dataset is a set of points of culture medium composition associated with a positive or negative observation of growth. This “binary data” is used in the literature of:

a) manual curation of metabolic models, which are “greedy” methods in nature (Orth & Palsson, 2012; Bartell et al, 2017);

b) automated refining algorithms, which have been historically “greedy” (Loira, 2012; Devoid et al, 2013) and recently have had the first “non-greedy” examples (Hartleb et al, 2016; Fritzemeier et al, 2017; Hartleb et al, 2018). Binary data was also used to assess CarveMe, a “top-down” reconstruction algorithm (Machado et al, 2018);

c) validation or refining using an analysis of gene essentiality (Henry et al, 2010), whose very definition uses this binary growth/non-growth concept (Rancati et al, 2017);

d) validation via Phenotype MicroArrays, which is quickly becoming the standard because it is a high-throughput technology (Blumenstein et al, 2015; BIOLOG, 2018).

A “greedy” optimization selects a local optimal at every step and may be trapped in a local optimum, whereas a “non-greedy” is global but usually more complex. A “top-down” draft reconstruction algorithm starts from general model skeletons and tailors them down to the specific data input, whereas a "bottom-up" builds the draft up from that input. CarveMe is likely the first top-down algorithm implemented in software.

The binary data is the most usual. The second most common is quantitative growth (a yield or a rate) under carbon limitation or no limitation at all (exponential growth). It cannot be generated by current high-throughput technologies like Phenotype MicroArrays™. Growth under these limitation regimes has been historically studied for animal parasites/symbiotes like *E. coli*, agricultural crops, environmental microbial communities and fermentations like beer and yoghurt. Thus, this kind of phenotypic data is hereby deemed "legacy data".

In fact, the classic Monod model is a relationship between specific growth rate $\mu$ ($h^{-1}$) and the concentration of a single limiting substrate $S$ ($g\ L^{-1}$), and in its original publication it was validated with legacy data of *E. coli*, *Mycobacterium tuberculosis* and *Bacillus subtilis* growing under carbon limitation or no limitation (Monod, 1949).

More comprehensive phenotypic datasets are rarer and found mostly on specific fields. For example, those including products other than biomass, like $CO_2$, acetate, ethanol, etc, are found in the literature of bioprocesses, bioproducts and specific organisms (Groot et al, 1992). Datasets in limitation regimes involving substrates other than carbon are found in agronomy (De Wit, 1994) and microbiology (Egli, 1991; Zinn et al, 2004). This kind of phenotypic data belongs to a general superclass hereby named "*N*-data" such that *N* is the number of growth-limiting factors in the dataset. Then, "legacy data" is "*1*-data" and "binary data" is "*0*-data".

A common pattern for *N*-datasets of $N > 0$ has been reported for a variety of organisms and sets of *N* substrates (Egli, 1991; Zinn et al, 2004; Taciro, 2008). The literature has reviews on cell growth models (Kovárová-Kovar, 1998; Carcano, 2010) and attempts to unify them (Nijland et al, 2008). Although these studies are based mostly on fitness to empirical data, theory has also been developed (Gorban et al, 2011; Gorban et al, 2016). Three common general metabolic principles were identified:

I. First kinetic principle: the relationship between cell growth rate and substrate availability is best described by a saturation type of curve;

II. Continuity and smoothness: not one but multiple factors concomitantly impact the growth rate, and, moreover, the transition between growth regimes (set of impacting factors and how they impact growth) is continuous and smooth;

III. Multiple Objective (MO): organisms maximize their growth rate and also the consumption and storage of the most critical factors, in order to hinder competition.

Principle I was proposed in 1912 and has been widely accepted. The Monod model is an example, because it is indeed a saturation type of curve. Principle II is tacitly assumed in all these reviewed studies, but it can be demonstrated by absurd as shown in section 5.1.1.

That study from 2011 focuses precisely in Principle III and concludes “We avoid any kinetic modeling. Nevertheless, adaptation is a process in time. We have to create a system of dynamical models”. Such “kinetic modeling” is the object of the follow-up work from 2016.

The “kinetic” and “stoichiometric” aspects of phenotypic *N*-data are defined in the study that generalizes the pattern observed for $N > 0$ (Zinn et al, 2004). The first is “the control of the cell growth rate by two nutrients at the same time” and the second is “the restriction of the amount of biomass by two nutrients at the same time”. It is also stated:

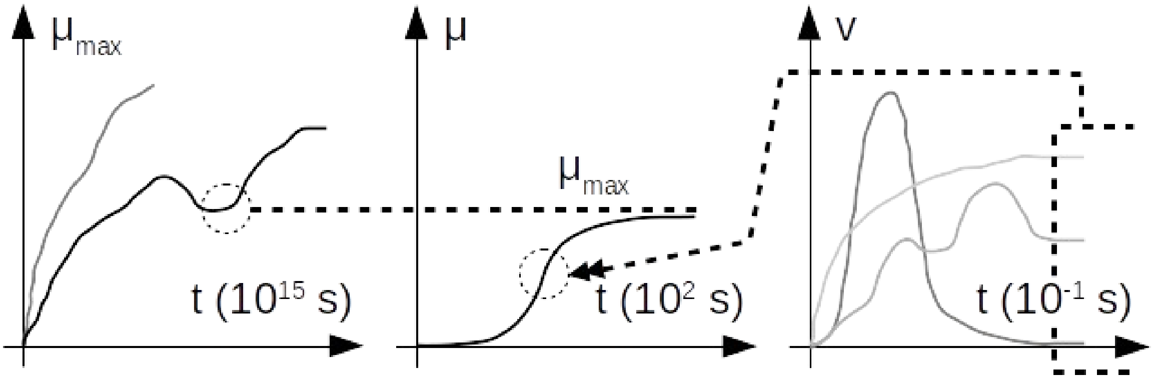
> The two aspects cannot be linked in practice, because kinetics deal with the growth rate as a function of the nutrient concentration in the culture broth (“growth rate limitation”), whereas the stoichiometric aspect specifies the relationship between biomass concentration and the therefore consumed nutrients (“biomass amount limitation”). (Zinn et al, 2004, p. 265)

However, if kinetics cannot be linked to stoichiometry in practice, then the final stoichiometry observed in a microbial population growing in steady-state (a chemostat culture) cannot be predicted from the kinetic path taken by this same population in the previous transient-state (continuous culture before reaching the steady-state). This is intuitively absurd and may be the result of not separating between three time-frames.

The short-term time-frame is a fast, effective and highly entropic individual response to intense environmental changes like a sudden pulse of carbon source. The transient-state lasts few seconds (Van Heerden et al, 2014). Then the metabolism reaches a relatively steady-state with much slower variations (minutes to hours), that is, a metabolic pseudo-steady-state with relatively constant internal concentrations and apparent global growth stoichiometry.

The medium-term time-frame is a slow, mildly effective and mildly entropic collective response to mild environmental changes such as the relatively slow depletion of substrates and accumulation of products due to the growth of the population itself. The effectiveness is decreased by subpopulations, bi-stability and mutations over some generations, meaning that a part of the population responds badly to the changes. The better the fitness of the species, the larger this part is relative to the population (Gorban et al, 2011). The transient state lasts from minutes (bacterial cultures) to days (animal cell cultures) or even years (spores in dormant state). The medium-term transient-state is the short-term steady-state. In a recent work, these are called respectively “time-local” and “time-global” steady-states (Reimers & Reimers, 2016);

The long-term time-frame is a very slow, binarily effective and negatively entropic species response to very slow changes such as the shift from anaerobic to aerobic atmosphere on Earth caused by life itself. The binary effectiveness is extinction or prevalence. Negative entropy is because the meaningful information content in the DNA increases from random processes, at a great expense of energy during many life-cycles (Adami, 2012). Thus, the long-term transient-state is the medium-term steady-state, which is another link between the “kinetic” and “stoichiometric” aspects of growth. The long-term steady-state is the biosphere in thermodynamic equilibrium. The time-frames are compared in Figure 1.



a) species time-frame b) population time-frame c) individual time-frame

*Figure 1: The three time-frames considered for transient-state kinetics and steady-state stoichiometry (μ is specific growth rate and v reaction rate or metabolic flux). In species time frame (a) the grey species was extinct and the black prevailed. Each point of the transient-state of the population time-frame (b) corresponds to a steady-state flux distribution in individual time-frame (c).*

In Figure 1, each point of transient-state in a time-frame is a point of steady-state in the immediately shorter time-frame. This arises from Principle II and is a way to link kinetics to the stoichiometry of growth.

## 3.3 On the genome-scale metabolic models

The most general format of a metabolic model is a system of differential equations expressed in vector form as $\vec{m}'(t)=\lambda\cdot\vec{r}[\vec{m}(t)]$, where $\vec{m}(t)$ is the column-vector of all masses of metabolites, $\vec{m}'(t)$ its derivative with respect to time *t,* $\vec{r}$ is the column-vector of reaction rates (with a kinetic submodel for each reaction) and $\lambda$ is the stoichiometric matrix with metabolites as rows and reactions as columns. This is a kinetic metabolic model.

The steady-state turns it into a stoichiometric model: the derivative term $\vec{m}'(t)$ is set to zero and the kinetic submodels are eliminated from $\vec{r}$, which becomes a simple vector of fluxes $\vec{v}$, yielding the common formulation $\vec{0}=\lambda\cdot\vec{v}$ as exemplified in Figure 2.

a) set of reactions

$$\begin{aligned} A[e] &\rightarrow A \\ A + 2B &\rightarrow C \\ C + 3A &\rightarrow D + B \\ D &\rightarrow D[e] \end{aligned}$$

b) stoichiometric matrix

$$\lambda = \begin{bmatrix} +1 & -1 & -3 & 0 \\ 0 & -2 & +1 & 0 \\ 0 & +1 & -1 & 0 \\ 0 & 0 & +1 & -1 \end{bmatrix}$$

c) Kinetic

$$\begin{bmatrix} A'(t) \\ B'(t) \\ C'(t) \\ D'(t) \end{bmatrix} = \lambda \cdot \begin{bmatrix} v_1(t) \\ k_2\ A(t)\ [B(t)]^2 \\ k_3\ C(t)\ [A(t)]^3 \\ v_4(t) \end{bmatrix}$$

d) Stoichiometric

$$\begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \end{bmatrix} = \lambda \cdot \begin{bmatrix} v_1 \\ v_2 \\ v_3 \\ v_4 \end{bmatrix}$$

*Figure 2: Example of kinetic and stoichiometric models for the same set of chemical reactions. The "[e]" suffix is to denote external metabolites, that is, those out of the control volume (cell) and that don't need to be balanced. Reactions 1 and 4 are exchange reactions.*

As mentioned in section 1, genome-scale models may have thousands of reactions. A kinetic model of that size is impractical since each reaction can have a different kinetic submodel with its set of parameters (*k* and exponents) and non-linear dependence on the coordinates of $\vec{m}(t)$. Besides, a genome-scale kinetic model is unnecessary if the interest is in the population time-frame (Figure 1), which usually is the case of bioprocess R&D.

Thus, the stoichiometric model is preferred. As shown in Figure 2, this kind of metabolic model is merely the stoichiometric matrix (hence the name). This matrix is a compact mathematical representation of the metabolic network.

### 3.3.1 Determining the genome-scale metabolic network

Currently, the standard method of determining a genome-scale metabolic network for an organism is to input the assembled genome in an automated draft reconstruction software, manually refine the resulting draft and then perform simulations with the refined network to validate it against phenotypic data (Ferrarini et al, 2015; Faria et al, 2018).

The vast majority of the draft reconstruction softwares compare the gene sequences of the input genome to databases of Gene-Protein-Reaction associations (GPR) to find gene orthologs, and then select the corresponding reactions above a certain threshold. This is a bottom-up approach because the network is built from parts of it (reactions). Today, the most popular fully automated draft reconstruction platform is the modelSEED webservice, which has an interface – also via web – with extra features named KBase.

In a recent example of the whole pipeline of network construction (Chatterjee et al, 2017), the tool used to obtain the draft was PRIAM (Claudel-Renard et al, 2003). PRIAM is a database updated regularly with software simple to use which compares gene sequences by position-specific score-matrices. However, the reactions are identified by Enzyme Commission numbers (EC), which can be ambiguous: a generic EC number may be associated to many reactions, of which only a few were actually detected in the genome.

CarveMe is possibly the first software that takes the opposite top-down approach, in which the network is built by tailoring a preset network. It has skeleton networks for each class of organism and a database of gene sequences which is used first to identify which skeleton is the best starting point and then to tailor it with the particularities found in the input genome, such as lack of a specific reaction (Machado et al, 2018).

The gene orthology analysis is usually integrated in the reconstruction software, like in the PRIAM, modelSEED and CarveMe examples. In the next step, the refinement, some solutions use this information as input to orient the process. For example, calculate scores based on this information for each GPR in the database and prioritize adding the reactions with higher score. There are several software and criteria for the gene orthology analysis.

The standard is a comparison using the Basic Local Alignment Search Tool (BLAST) with a setting of an *e*-value, which is an error tolerance, so the lower the more strict (Wallner et al, 2004). A recent review discusses more than 600 articles on techniques and computational tools published between 2011 and 2017 (Nichio et al, 2017). This large amount of publications shows this is an intense subject of current research, what also means there are still many problems yet to be solved, including the computational speed of the tools and their learning curves (Cosentino & Iwasaki, 2019).

As mentioned in section 3.2.1, the refining process can be manual but there have been many recent efforts to automatize it with many different approaches. The first review article referenced in this section states:

> One significant downside of early methods was that they typically produced nonfunctional models that were incapable of simulating biomass production. Significant additional manual curation was required just to enable biomass production. (...)
>
> Integrated gap filling is one of the most important features in a reconstruction, as manual efforts at gap filling can be extremely time-consuming. When performed manually, one has to identify the gaps and candidate reactions to complete a pathway/network. (...)
>
> Automated gap-filling solutions still require manual inspection for further refinement, as reactions can be arbitrarily added to restore model connectivity and pathway completeness. This fact calls for the continuous development of improved gap-filling algorithms. To address that issue, all platforms that integrate gap filling usually provide their own algorithms. Those algorithms are variations of the GapFill algorithm. (...)
>
> One of the drawbacks of GapFill and its variations is the use of mixed-integer linear programing (MILP) to determine the minimum set of reactions to be added to the model. MILP can be difficult to solve, particularly if one lacks commercial optimization software (e.g., CPLEX). In one example case, GapFill was found to take over 14 h to find an optimal solution for a single model of a prokaryote. (...)
>
> Recently, methods have been introduced that use linear programing (LP) to substantially reduce the computation time required for gap filling. (…) FastGapFilling [40] also utilizes an LP formulation (…) performing up to three times faster when compared with a MILP formulation. (Faria et al, 2018)

The idea of the GapFill algorithm (Kumar et al, 2007) is first to find the gaps and then try to solve them either by reversing existing reactions or by adding new reactions from an external database which includes exchange reactions (import or export a metabolite). Each one of these two steps (finding and solving) is a different Mixed-Integer Linear Programming optimization problem (MILP). The first step is to find which metabolites cannot be produced in the network, that is, which have gaps in all their production routes (Figure 3).

Given a draft network with:

| | |
|---|---|
| $\lambda$ | Constant stoichiometric matrix (metabolites indexed by i and reactions by j) |
| $\vec{v}, \vec{U}, \vec{L}$ | Variable set of fluxes + constant upper and lower bounds (all indexed by j) |

Define two arbitrary tolerances to encode "OR" constraints and necessary bounds:

| | |
|---|---|
| $\epsilon_{min}$ | Minimum flux through a metabolite |
| $\epsilon_{max}$ | Maximum flux through a metabolite |

Define 2 sets of binary variables:

$$x_i \quad \begin{cases} 1 & \text{if metabolite i can be produced} \\ 0 & \text{otherwise} \end{cases}$$

$$w_{i,j} \quad \begin{cases} 1 & \text{if reaction j that produces metabolite i is active} \\ 0 & \text{otherwise} \end{cases}$$

| | | |
|---|---|---|
| Problem: | $max \sum_i x_i$ | , subject to: |
| i) Irrev. reactions (subset of M), $\lambda_{i,j}>0$ : | | $\lambda_{i,j} v_j \geq \epsilon_{min} w_{i,j}$ |
| ii) Irrev. reactions (subset of M), $\lambda_{i,j}>0$ : | | $\lambda_{i,j} v_j \leq \epsilon_{max} w_{i,j}$ |
| iii) Rev. reactions (subset of M), $\lambda_{i,j} \neq 0$ : | | $\lambda_{i,j} v_j \geq \epsilon_{min} - \epsilon_{max} (1 - w_{i,j})$ |
| iv) Rev. reactions (subset of M), $\lambda_{i,j} \neq 0$ : | | $\lambda_{i,j} v_j \leq \epsilon_{max} w_{i,j}$ |

Metabolites which can be produced have at least 1 production route:

| | |
|---|---|
| v) Irrev. reactions (subset of M), $\lambda_{i,j}>0$ :<br>vi) Rev. reactions (subset of M), $\lambda_{i,j} \neq 0$ : | $\sum_j w_{i,j} \geq x_i$ |

The usual constraints of metabolic network linear optimization:

| | |
|---|---|
| vii) The input bounds for each flux (for all j): | $L_j \leq v_j \leq U_j$ |
| viii) Steady-state + exchange reactions (for all i): | $\sum_j \lambda_{i,j} v_j \geq 0 \iff \lambda \cdot \vec{v} \geq 0$ |

*Figure 3: The first step of the GapFill algorithm for single-compartment networks is a MILP maximization problem subject to 8 sets of restrictions (i to viii). Adapted from the code supplied with Kumar et al, 2007.*

As shown in Figure 3, GapFill needs one binary variable for each metabolite and another for each reaction that produces it. Thus, reactants in irreversible reactions do not need to be considered and they are indeed excluded from the subsets in restrictions i and ii. For reversible reactions, all participating metabolites generate binary variables. The restrictions iii and iv are a common way to encode constraints of two disjoint intervals. For example, setting $\epsilon_{max}=100 \gg \epsilon_{min}=0.01$ , an active reaction (*$w_{i,j}$ = 1*) must satisfy $0.01 \leq \lambda_{i,j} v_j \leq 100$ , which is another way to state an "OR" condition: $\{\lambda_{i,j}>0 \wedge v_j>0\} \vee \{\lambda_{i,j}<0 \wedge v_j<0\}$ . The computing time to solve MILP problems depends mostly on the total number of integer variables.

The result of this first step is the list of metabolites whose *$x_i$ = 0*. The second step is a very similar problem. A new set of binary variables is added, *$y_k$*, which is 1 if the *k*-th reaction from the external database was added to the draft network, and the objective function is changed to minimize the sum of these *$y_k$*. All the reactions of the external database are reversible, so constraint sets i and ii are eliminated, whereas the reaction subsets of iii and iv are changed to the database.

While GapFill is formulated as two sequential MILP problems, FastGapFilling is iterative (Latendresse, 2014). The draft network and the external reaction database are merged into one supernetwork, whose biomass minus the penalties for adding reactions from the database are maximized: $\delta \cdot v_{bio} - \sum_{j \in \text{database}} c_j \cdot v_j$ .

What changes in every iteration is the value of the weight $\delta$ . It starts as the number of reactions in the external database and is updated after each maximization. When a valid solution is found ($v_{bio} > 0$), $\delta$ is decreased and otherwise it is increased. The value of $\delta$ in the next iteration is bounded by the current value: $0.5\,\delta_k \leq \delta_{k+1} \leq 1.5\,\delta_k$ . The stop criterion is a function of the values of this variable weight.

This is a greedy algorithm and is susceptible to being trapped in a local optimum, e.g. when the increase in the variable weight is not enough for the linear solver to find a vertex of the polytope representing the linear problem which is better than the one corresponding to the last valid solution. The review article has a mistake: FastGapFilling is not up to three times faster than a MILP formulation but up to three orders of magnitude faster.

The validation of a metabolic network with phenotypic data can be considered the last step of the process of determining the structure of network, since this structure must be able to explain what is observed in experiments. This validation is some kind of simulation of the network with constraints representing the phenotypic data.

The standard basic simulation technique for genome-scale metabolic networks is Flux Balance Analysis (FBA) (Orth et al, 2010), which is a linear optimization and is implemented in many software solutions readily available (including webservices such as KBase). In fact, the linear problem solved in every iteration of FastGapFilling is an FBA simulation.

If constraints representing a point from a phenotypic dataset are added to this problem, FastGapFilling would be extended from resolving only connectivity problems to also resolving wrong predictions. That is to merge the refining and validation steps, what could be more efficient than doing them sequentially. The merged approach is used in many algorithms, including the gap-filling module of KBase, which allows the input of a culture media composition to gap-fill against.

However, most of the implementations of such merged approach can only validate 1 phenotypic point at a time, due to the greedy nature of the underlying algorithm. For example, in the hypothetical example of extending FastGapFilling, if the phenotypic dataset has two points with distinct values of glucose uptake, only one can be set at a time for the FBA.

This limitation is the main motivation to develop non-greedy algorithms able to validate multiple points at once, since this may generate an optimal solution that is better than the union of all refined networks generated for one point at a time (Hartleb et al, 2016). One of such non-greedy refining/validation algorithms is GlobalFit. It minimizes the global set of changes in the network in order to minimize the number of wrong predictions among a binary phenotypic dataset. It is thus a bi-level problem:

- inner problem (linear, FBA): maximize the biomass flux $v_{bio,k}$ of the $k$-th simulation in order to compare it to the viability threshold $T_{g,k}$ or $T_{h,k}$ of the $k$-th phenotypic binary point. The wrong predictions are those with $v_{bio,k} < T_{g,k}$ if $k$ is a growth point and $v_{bio,k} > T_{h,k}$ if $k$ is a non-growth point;
- outer problem (mixed-integer linear): minimize the sum of the penalties associated to the changes in the model. The penalties are constants predefined for each type of change, which is represented as a binary variable (changed/unchanged).

The allowed changes are: (i) removals or (ii) reversibility change of existing reactions; (iii) additions of reactions from the list of potential reactions; (iv) removals or (v) additions of metabolites in the biomass reaction. So the most important inputs for GlobalFit are:

- 1 draft network + 2 lists of reaction removal penalties (forward and reverse) + 1 list of allowed reaction reversals (i and ii);
- a list of potential reactions + a list of penalties for adding each one (iii);
- 1 list of allowed metabolite removals + 1 list of allowed metabolite additions + 1 single value for the stoichiometric coefficient of any metabolites added (iv and v);
- 2 list of phenotypic points (“on” and “off”). Inside each of these lists, each $k$-th point has its own list of influxes, an optional list of reactions to remove (knock-out) and a value for its $T_{g,k}$ (“on”) or $T_{h,k}$ (“off”).

### 3.3.2 Maintenance energy submodels

"Maintenance energy (requirement)" does not have a clear and unambiguous definition despite being a very important concept implemented in most metabolic models. It is the notion that part of the energy harvested by an organism is not stored nor spent on growth but rather on vital functions which consume energy uninterruptedly or else the organism dies.

It is observed empirically and directly as biomass consuming itself to stay alive when no other energy sources are available. For example, production of $CO_2$ in a culture where the mass of cells is decreasing and there is no carbon source substrate available. It is also observed indirectly in a plot of specific growth rate ($\mu$) in function of energy source uptake rate (e.g. $q_{glucose}$), by extrapolating the points towards the origin; the line intercepts *(m, 0)*, with $m > 0$, so cells consume energy even when they are not reproducing.

The second method is indirect because the points *(q, $\mu$)* are measured in chemostat cultures, which cannot reach $\mu = 0$. A chemostat culture is by definition in a dynamic steady-state, condition in which the material balance of biomass implies $\mu = D = F/V$, where $D$ is dilution rate, F is the net inlet flow rate and $V$ is volume (Ziv et al, 2013). But $\mu = D = 0 \Leftrightarrow F = 0$, what means substrates are not replenished and deplete while products are not washed-out and accumulate, contradicting the premise of steady-state.

The simplest and most traditional maintenance energy models are Herbert's and Pirt's, which respectively correspond to the direct and indirect observations of the phenomenon. These models were conceived empirically in the 1960's and the theory associated with them is still object of debate (Van Bodegom, 2007; Wang & Post, 2012). It is now known that these two models are related by a linear transformation on their parameters (Equations 1 to 6)

Herbert 1: $\mu(S) = \mu_{max,H} \cdot S/(K_S + S) - a_H$ (1)

Herbert 2: $q(S) = \mu_{max,H} \cdot S/(K_S + S) \cdot 1/Y_G$ (2)

Pirt 1: $\mu(S) = \mu_{max,P} \cdot S/(K_S + S)$ (3)

Pirt 2: $q(S) = \mu_{max,P} \cdot S/(K_S + S) \cdot 1/Y_G + m$ (4)

Relation 1: $a_H = Y_G \cdot m$ (5)

Relation 2: $\mu_{max,P} = \mu_{max,H} - a_H$ (6)

The function $g(S) = S/(K_S + S)$ in Equations 1 to 4 is the Monod model and $S$ is the concentration of the limiting energy-source substrate (carbon, for chemoorganotrophs). In theory, it could be any function that satisfies $g(S >> K_S) \rightarrow 1$ (Principle I). The ideal biomass yield from substrate $S$ under constant volume is $Y_G = Y_{X/S}{}^* = dX^*/dS^* = \mu(S)^*/q(S)^*$, where the asterisk means ideal, that is, without maintenance energy ($m = a_H = 0$). In this sense, the energy spent on maintenance is conceived as an unavoidable minimum waste.

Therefore, these models are fundamentally of kinetic nature. But a metabolic network is of stoichiometric nature, in steady-state. The most usual way to link those two in order to a include maintenance submodel in a metabolic network is to add an ATP sink reaction and constrain its flux (both lower and upper bounds to the same non-zero value). An ATP sink is an imbalanced reaction which converts ATP to ADP without producing anything. This reaction (or equivalents, such as an external proton sink) is found in most published genome-scale metabolic networks (Broddrick, 2018).

The value of the flux in the ATP sink is not directly related to observable variables such as $m$ or $a_H$, precisely because there are many possible pathways in the metabolic network that convert between the measured metabolites with different yields. For example, from glucose to biomass. There is, however, a path of maximum yield $Y_{X/S}{}^*$ allowed by the network stoichiometry, activated in an FBA simulation which maximizes $v_{bio}$ - $v_{glucose}$. This yield should be higher than that obtained experimentally and it can be brought down to the correct value by increasing the flux of the ATP sink.

Just like the Herbert and Pirt models, this procedure also implicitly assumes that the culture is under energy limitation. This is because there is no reason to believe $1/Y_G$ and $m$ (or $a_H$) are constant unless they are at minimum possible values, which is a reasonable assumption under energy limitation. Under energy excess this may not be the selected phenotype over the course of evolution, because competitors may grow faster using a less efficient metabolism, what is indeed observed (Szenk et al, 2017).

Even under energy limitation there is also the bias in the flux distributions despite individually accurate predictions for the measured fluxes, since there is no ATP sink in reality. Regardless of what maintenance really means, the energy spent on it is scattered all over the metabolism in futile cycles, e.g. erroneous synthesis of DNA, degradation of the erroneous part and re-synthesis of it. And these cycles are not equivalent to ATP sinks because not all of their metabolites are completely balanced, so they have net consumption and production of metabolites which in reality are compensated by the rest of the network.

Thus, in the condition of energy excess not only these two maintenance models themselves are hardly accurate as the bias in the flux distribution is probably worse since more energy to dissipate means more futile cycle possibilities. This motivated the pursue of a more general maintenance model to improve on the condition of energy excess, of which the most traditional example is Extended Pirt or Epirt (Pirt, 1982).

Extended Pirt replaces the $m$ parameter by a linear function $m_1 + m' (1 - \mu/\mu_{max})$. Naturally, the same dataset fit with Pirt and EPirt will not produce $m = m_1$, but instead $m$ will be an intermediary value assumed by the linear function of EPirt. Even though this simple extension may fit the available data, as it did in the publication from 1982, there is no theoretical reason to believe the maintenance energy varies linearly with $\mu$.

In fact, there is data that shows it does not, as Pirt found himself (Pirt, 1987): for low values of $\mu$, like $\mu < 0.1\ \mu_{max}$, all these three linear models (Herbert, Pirt and EPirt) seem to overestimate the maintenance energy requirement. In 1987, Pirt proposed a model based on the assumption that at low $\mu$ a fraction of the population is in a dormant state consuming much less energy for maintenance than what would be expected. In other words, at low $\mu$ the common hypothesis of uniform population introduces significant error and the average $\mu$ is not a good representation of the physiology of the dormant subpopulation.

Recently, a phenomenological model for the maintenance energy requirement has been proposed (Fernandez-de-Cossio-Diaz & Vazquez, 2018). Recognizing that to this day "(…) we are lacking a theory explaining the maintenance energy demand", the authors derive a complex physical model based on the idea that most of this energetic demand is for molecular motors that fluidize the cytoplasm in order to counteract molecular crowding. They state that cytoplasm is a gel-like structure, very different from an ideal solution, and molecular crowding impedes vital processes like protein synthesis.

This recent model is designed for eukaryotic cells, highly compartmented and with high maintenance energetic costs, but some data of *E. coli* suggest that it might apply to prokaryotic cells as well. No examples of application of this model or simplified versions of it inside steady-state metabolic networks was found.

## 3.4 On the *in silico* experiments

Most simulation techniques currently used in metabolic networks fall under one of these three categories: structural, constraint-based or hybrid between them. The structural class is focused on the topological structure of the network as a whole, that is, the connectivity between metabolites, the possible paths to consume or produce sets of metabolites, minimal precursor sets, etc. The constraint-based class is based on constraining the metabolic network and simulate in detail a subset of all the possibilities of it.

Thus, structural analysis is bottom-up because the overall behaviour of the network is predicted from its constituents, whereas constraint-based analysis is top-down because external constraints are used to predict a more specific behaviour. The first is less biased but also in general much more computationally complex than the second. Hybrid techniques are used to combine the advantages of both.

The earliest and simplest techniques of the two complementary classes began being developed in the 1980's: Elementary Mode Flux Analysis (EFMA) (Schuster et al, 1999; Zanghellini et al, 2013) and Flux Balance Analysis (FBA) (Fell & Small, 1986; Orth et al, 2010). These are the two basic frameworks, respectively bottom-up and top-down.

A "mode" is a flux distribution in which all metabolites are balanced (steady-state) and all irreversible reactions are in the appropriate direction (thermodynamics). EFMA is to determine a minimal set of "elementary modes" which span the whole set of modes and FBA is to search for one specific mode that optimizes a linear objective function.

An elementary mode is a set of reactions that together perform a certain task whereas a "minimal cut set" is a set of reactions that must be deactivated to prevent a given task. It has been shown that: most network consistency problems like finding blocked reactions are easy to solve, finding an elementary mode is easy but finding a specific one is not and finding a minimum (not minimal) cut set is hard (Acuña et al, 2009).

A "minimal precursor set" is a set of metabolites that an organism may obtain from the environment and which enables it to produce a set of metabolites of interest. Software tools are available to find such sets in a purely topological analysis, without considering the stoichiometry (Acuña et al, 2012a), and also including it (Andrade et al, 2016).

Furthermore, advanced graph theory has been developed to compute “stories”, which are possible scenarios explaining the flow of matter through a metabolic network given a list of metabolite concentration changes and a network (Acuña et al, 2012b) and that was applied on a case-study of yeast response to cadmium exposure (Milreu et al, 2014).

On the other hand, constraint-based modelling has also been object of study. The most basic technique after FBA is Flux Variability Analysis (FVA), which is to determine the minimum and maximum admissible fluxes through each reaction of the network. This is simply a set of sequential FBA simulations, two for each reaction. FVA allows evaluation of the whole set of admissible flux distributions, identification of blocked reactions (Yousofshani et al, 2013), and thermodynamically inconsistent cycles (Müller & Bockmayr, 2013).

Parsimonious enzyme usage FBA (pFBA) is a bi-level linear optimization: the inner problem is to maximize the growth rate and the outer problem is to minimize the sum of all fluxes of gene-associated reactions. The underlying assumption is that selective pressure favours the fastest growing organisms, which in turn are likely those with the most stoichiometrically efficient enzymes. The inner problem is solved first by ordinary FBA, then the biomass flux is constrained to between 90% and 100% of its maximum and finally FVA is carried-out to select the flux distributions with the lowest sums of fluxes (Lewis et al, 2010).

Multi-Objective (MO) optimizations like pFBA forms an entire subclass of constraint-based modelling. They are particularly interesting given the MO nature of evolutionary systems, what is summarized by Principle III of section 3.2.1. A recent work briefly compared the three most common methods of MO optimization and proposed MO-FBA and MO-FVA implementations based on one of them, “Benson type” algorithms (Budinich et al, 2017).

The three methods discussed are, in order of complexity: optimize an objective function that is a weighted sum of all the multiple objectives (used globally by FastGapFilling); keep only one objective function and introduce the others as constraints (used by pFBA and GlobalFit); and approximate the outer shape of the solution space by a convex polytope to identify vertices of it (“Benson type”, used by MO-FBA and MO-FVA; used by FastGapFilling inside each iteration).

Another subclass of FBA modifications is focused on the simulation of dynamic systems. The earliest is Dynamic FBA (dFBA), which “places the steady-state constraint-based formulism inside a discrete time step dynamic approximation that uses Michaelis-

Menten kinetics to simulate nutrient uptake" and has been recently used to simulate microbial community dynamics in genome-scale (Mellbye et al, 2018). This is a recent example of application, but dFBA was first described in 2002 (Mahadevan, 2002), and it is based on the assumption that the transient-states of a long time-frame are steady-states of a shorter time-frame (illustrated in Figure 1).

Unsteady-state FBA (uFBA) is another example of this kind, based on this same assumption but also including an intermediate step of Principal Component Analysis (PCA) to determine which are the most dissimilar time intervals in a phenotypic dataset. One metabolic model is used for each time cluster (Bordbar et al, 2017).

As to the hybrid techniques, the best conceptual example is Elementary Flux Vectors (EFV), because it is EFMA extended with linear constraints, which are the essence of FBA (Klamt et al, 2017). A recent work from the same research group expands on the theory and applies it in an example of metabolic engineering (Klamt et al, 2018).

An important result of the latter work is to demonstrate that rate-optimal solutions such as the ones generated by most constraint-based methods are only sometimes similar to yield-optimal solutions, although yields are usually the aim of rational strain design. The authors have also discussed methods to compute projections of the whole yield-space of a genome-scale metabolic network into 2 or 3 selected yields.

Despite being the best conceptual example given the two frameworks that became standards, EFV is not the earliest hybrid technique. Some earlier examples are:

a) Flux Coupling Analysis (FCA) (Burgard et al, 2004), whose core concept was incorporated into pFBA to group reactions into 5 classes;

b) Macroscopic Biorreaction Models Analysis (MBMA) (Provost et al, 2006), which selects modes from EFMA (not by optimization like FBA does) and applies Principle II to model population dynamics (like dFBA is to FBA);

c) Hybrid Cybernetic Modelling (HCM) and Lumped-HCM, which are improvements over MBMA (Song et al, 2009; Song et al, 2013);

d) Dynamic Reduction of Unbalanced Metabolism (DRUM), which divides the metabolic network into subnetworks (artificial compartments) connected by metabolites whose concentration varies slowly (Principle II) and then applies MBMA in each subnetwork (Baroukh, 2014).

Another approach in hybrid techniques is the use of stochastic methods. For example, to predict dynamic behaviour from a specific starting condition (Gilbert et al, 2017) or to characterize all possible flux distributions of a metabolic network (Braunstein et al, 2017). The latter can be understood as computing a probability distribution for the possible values of each flux of the network, instead of the minimum and maximum values determined by FVA.

Stochastic methods can be very computationally intensive and impractical for genome-scale metabolic networks, since they are in reality brute-force approaches to perform complex simulations. The two examples mentioned here use approximations to tackle this issue. One important remark is that both apply Principle II, the first like in Figure 1 (section 3.2.1) and the second in the sense that for every flux of the network is represented by a continuous and smooth probability distribution function.

Yet another kind of hybrid analysis is the search for knock-out sets which optimize the production or consumption of a set of multiple metabolites. A knock-out is to completely remove a gene, which removes a set of reactions from the metabolism, constraining it. This is used to remove metabolic functions that compete with the task to be optimized. For example, a byproduct that consumes the same precursors as a bioproduct of interest. Thus, knock-out analyses are primarily structural and are related to the concept of minimal cut sets. Still, they can be improved with the addition of top-down constraints, yielding hybrid techniques.

GridProg is a recently published algorithm for this which implements a parsimonious optimization similar to pFBA in order to find knock-out sets for multiple metabolites (Tamura, 2018). Approaches like this can be used to evaluate the complementarity of different organisms with respect to their potential applications. However, such evaluation would be far from comprehensive, because not all competing functions can be removed without rendering the metabolism unviable. Fortunately, recent genetic engineering techniques allow down-regulating instead of completely removing a gene (Lv et al, 2015), but this possibility is not considered in any of the reviewed studies of this kind.

### 3.4.1 The case for a FBA-based approach

Previously, a small-scale model of *Pseudomonas* sp. LFM046 was proposed to explain the steady-state data and subject to Elementary Flux Mode Analysis (EFMA) (Taciro, 2008). A similar approach using a large-scale model of *Pseudomonas putida* KT2440 pinpointed a set of knock-outs (Poblete-Castro et al, 2013) which improved its $\mathrm{PHA}_{\mathrm{MCL}}$ accumulation to the level of the former strain (Poblete-Castro et al, 2014). EFMA is a structural analysis not scalable to genome-scale, so in these two examples some information was lost:

- part of the flux distribution was considered constant, but without knowing exactly which part was that and/or in which conditions it can be considered constant;
- elementary modes were analysed singly, what is blind to the small improvements still possible for the strain LFM046 (the large improvements are already in it);

In order to find optimization strategies for the production of PHA with the control of the monomeric composition, the metabolic model used for the *in silico* experiments must have the metabolisms of nitrogen and phosphorus. This because PHA accumulation is induced by limitation in these two nutrients and the PHA content reached is not the same for both.

However, including these metabolisms in the network of strain LFM046 made it too large for EFMA. The only solution for EFMA was to further constrain the network, for example by setting the proportion of biomass to PHA. But this defeats the whole purpose of structural analysis and is what constraint-based modelling is designed for.

As discussed in section 3.4, Flux Balance Analysis (FBA) is a simple yet powerful constraint-based modelling technique with many readily available software solutions, which are usually also capable of performing Flux Variability Analysis (FVA). Although FVA does not provide information on the probability of a flux reaching some value, it provides the admissible range of such value. FBA and FVA can also be used together to perform parsimonious enzyme usage FBA (pFBA), which can be more accurate than pure FBA.

Dynamic FBA is also present in many software solutions and this method was first published in 2002, being validated by its correct prediction of diauxic growth of *E. coli*. It is known that the strain *Pseudomonas* sp. LFM046 also exhibits an interesting pattern of diauxic growth when cultivated in fed-batch with a mixture of glucose and fructose: during the growth phase, when little to no polyhydroxyalkanoate (PHA) is detected, glucose is almost entirely depleted before fructose starts being consumed, but in the accumulation phase with both carbon sources replenished, the exact opposite is observed (data from Diniz et al, 2004).

Thus, there is strong experimental evidence not only of diauxic growth in this strain but also on a metabolic switch that completely inverses the diauxic pattern and that is related to the accumulation of PHA. Moreover, these are supported by independent evidence:

a) *Pseudomonas* sp. LFM046 has the genes of the fructose metabolization pathway via fructose-1-phosphate, which is equivalent to an Embden-Meyerhoff-Parnas pathway and was described for the similar organism *P. putida* KT2440 (Sudarsan et al, 2014). This pathway generates less reducing power equivalents per mole of carbon than the glucose metabolization pathway, which is the EDEMP cycle (Nikel et al, 2015);

b) *Pseudomonas* sp. LFM046 does not have the *gnd* gene (Cardinali-Rezende et al, 2015) and this was confirmed by enzymatic activity essays (unpublished). The enzyme encoded by *gnd* oxidises 6-phospho-gluconate producing reducing power equivalents. Also, it has the genes for the two following NAD/NADP transhydrogenases:

   - cytososolic: $\mathrm{NADPH} + \mathrm{NAD}^{+} \rightarrow \mathrm{NADP}^{+} + \mathrm{NADH}$ ,

   - membrane-bound: $\mathrm{NADH} + \mathrm{NADP}^{+} + \mathrm{H}^{+}_{\mathrm{external}} \rightarrow \mathrm{NAD}^{+} + \mathrm{NADPH} + \mathrm{H}^{+}_{\mathrm{internal}}$ ;

c) the deletion of the gene *gcd* in *P. putida* KT2440 made its PHA content and monomeric composition very similar to those of the strain LFM046, formerly named IPT046 (Poblete-Castro et al, 2014). This deletion blocked a periplasmatic pathway which oxidises glucose to gluconate while producing reducing power equivalents;

d) the prevalence of 10-carbon fatty-acid precursor reported in the structure of many rhamnolipids produced by *Pseudomonas* strains coincides with the prevalence of 10-carbon HA monomer in *P. putida* KT2440, *Pseudomonas* sp. LFM046 and even in a recombinant of the latter which should not exhibit that much $\mathrm{HA}_{\mathrm{MCL}}$. This precursor could be serving as a storage of reducing power excess (Cespedes et al, 2018);

e) the hypothesis of PHA as a sink of reducing power whose monomeric composition is a function of the ratio between the pool of reducing power equivalents and the pool of acetyl-CoA was confirmed in fed-batch experiments with *P. putida* strains.

Furthermore, a flux distribution was estimated from the measured fluxes for each instant using a dFBA-equivalent approach (Montano-Herrera et al, 2017).

Figure 4 represents the central metabolism of *Pseudomonas* sp. LFM046 in light of all these features in items a to e. The membrane-bound transhydrogenase (item b) is drawn next to the electron transport chain because it consumes proton-motive force.

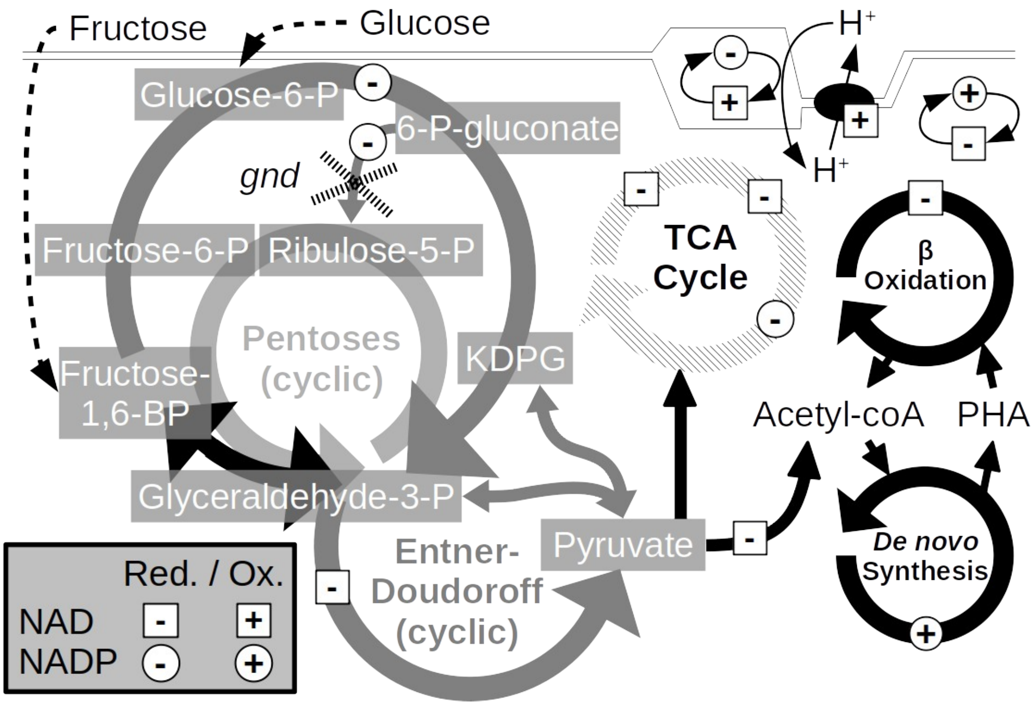


*Figure 4: Structure of the central metabolism of Pseudomonas sp. LFM046. The cyclic Pentose-Phosphate and Entner-Doudoroff pathways are the EDEMP supercycle (Nikel et al, 2015). On the right top corner, the electron transport chain oxidising NADH to pump protons out of the cell and the membrane-bound NADP/NAD transhydrogenase, which spends proton-motive force to convert NADH to NADPH.*

In item (e), the authors had measurements of 7 fluxes and used a simplified metabolic network which had only 6 degrees of freedom. They calculated a unique flux distribution for each point in time with 6 measured fluxes and used the extra one to estimate the error between measurements and predictions, with Metabolic Flux Analysis (MFA).

A genome-scale metabolic network has thousands of degrees of freedom, so this MFA approach is impractical. Instead, FBA and FVA can be used: for each point in time, constrain the measured fluxes, perform FVA, select a set of flux distributions whose temporal transitions are gradual and then further constrain the network such that FBA in each point produces the corresponding selected flux distribution. This is equivalent to dFBA using the same idea of pFBA to apply Principle II. If dynamic data is not available, an analogous approach can be used to ensure Principle II is satisfied between two distinct steady-states.

In Figure 4, the fatty-acids metabolism is represented as the two black cycles. The *de novo* synthesis pathway elongates a fatty-acid chain and the β-oxidation breaks it, both 2 carbons at each step, sequentially. For example, a precursor of 6 carbons can be formed either by elongating one of 4 carbons or breaking one of 8, but not directly from one of 2 or 7 carbons. The PHA monomers are de-routed from these two pathways.

In the *de novo* synthesis, the same set of enzymes anabolizes a $C_M$ precursor into a $C_{M+2}$ precursor, so this pathway cannot be regulated for each PHA monomer independently. Analogously, the β-oxidation also has one single set of enzymes for catabolizing $C_{M+2}$ to $C_M$. However, both pathways branch at the $C_{10}$ precursor because for $M \geq 12$ there are unsaturated precursors. Figure 5 depicts these cycles as spirals to show that.

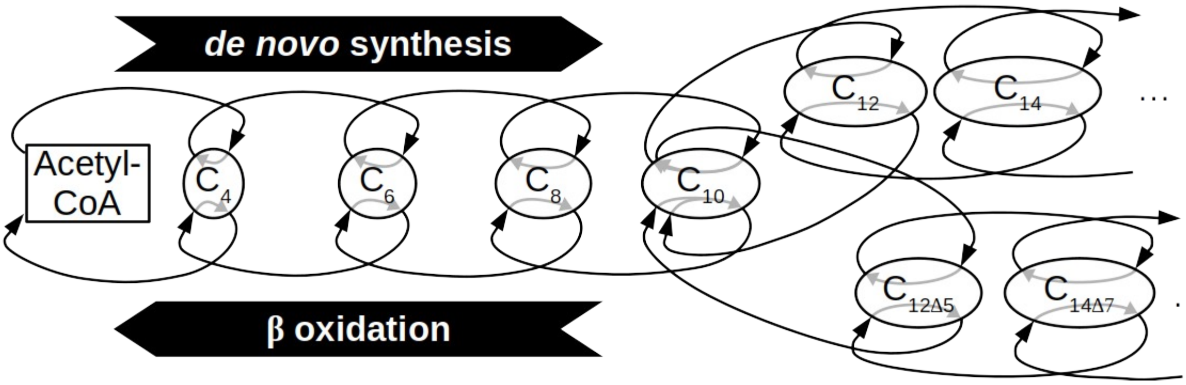


*Figure 5: Structure of the fatty-acids spirals: anabolism (de novo synthesis) and catabolism (β-oxidation). The $C_M$ metabolites are PHA monomers (3-hydroxyalkanoates) with M carbons. The branch on the bottom after $C_{10}$ is that of unsaturated monomers, where Δk represents a double-bond in the k-th carbon.*

The precursors $C_M$ with $M \geq 12$ are also the most abundant in biomass because they form the phospholipids of the cell membrane, whose most important properties like fluidity and selective permeability depend on the length and unsaturation degrees of the lipid chains (Li et al, 2018). In the $PHA_{MCL}$ of strains LFM046 and *P. putida* KT2440, the most prevalent monomer is $C_{10}$ and $C_{12\Delta5}$ is more abundant than $C_{12}$, suggesting that:

- the enzyme kinetics of both spirals is faster than that of the PHA polymerases, or else the cell would always accumulate much PHA, and not only when growth is slow;
- the enzymes that de-route the spirals into the unsaturated branches are even faster. The most important of them is a reversible isomerase (Mursula et al, 2001);

- unlike the revolutions within a spiral, the metabolism does distinguish between the two spirals as a whole and can control them separately. The enzyme sets are different and substrates too. The *de novo* synthesis uses acyl-carrier-proteins in place of the co-enzyme A and NADPH/NADP$^+$ instead of NADH/NAD$^+$.

A metabolic model of the strain LFM046 needs not to consider in detail the enzyme affinities with NAD(P) cofactors because of the NAD/NADP transhydrogenases (item b). In other words, even if the reactions of the fatty-acids spirals are set in the network with unrealistic ratios of NAD/NADP (Olavarria et al, 2015), the transhydrogenases will prevent infeasible simulations. If they form a futile cycle because of the energy expenditure in the membrane-bound reaction, that is unlikely to change significantly the predictions of the model because the maintenance energy already accounts for this kind of futile cycle.

These hypotheses on the spirals kinetics agree with the PHA composition produced by a recombinant strain constructed from LFM046 (item d). The native PHA polymerase was inactivated via UV and one from *Aeromonas* sp. was introduced. The former has low affinity for SCL monomers ($M < 6$) and the latter has low affinity for MCL monomers ($6 \leq M \leq 14$).

The expected result was a low or zero fraction of MCL monomers, but $C_{10}$ persisted, increasing from 0 to 7.6 %mol over the time of fed-batch cultivation, and also $C_{10}$ was always more abundant than $C_8$. Moreover, there was multiple evidence that the native polymerase was indeed inactive, so that was the first report of that SCL polymerase from Aeromonas sp. exhibiting significant activity on MCL monomers.

Hindered growth slows down the consumption only of the precursors $C_M$ with $M \geq 12$, so they build up until a chemical equilibrium is reached by slowing down the fatty-acids *de novo* spiral after $C_{10}$. A polymerase adapted for $C_{10}$ would then consume the $C_{10}$ and it would not build up, so before $C_{10}$ the spiral would continue fast and the PHA content in the cell would be high.

But a polymerase adapted for $C_4$ and $C_6$ would not consume $C_8$, which would form $C_{10}$, which in turn would build up until eventually the polymerization of $C_{10}$ is forced. This agrees with the higher $C_{10}$ content than $C_8$ and also the higher $CO_2$ relative production of the recombinant strain (less energetic efficiency).

Several reports in literature indicate a relationship between the maximum PHA content achieved by a strain and the average carbon oxidation state of their PHA (Table 3.1). This suggests high-content PHA accumulation is not limited by the supply rates of PHA precursors, which then continue in the fatty-acids *de novo* synthesis and accumulate as fat or equivalents.

*Table 3.1: Commercialized PHA (adapted from Możejko-Ciesielska & Kiewisz, 2016)*

| Strain | DNA manipulation | Carbon source | PHA type | Content (% wt) |
|---|---|---|---|---|
| *Alcaligenes latus* | - | sucrose | PHB | 75 |
| *Bacillus sp.* | - | sugar cane | PHB | 90 |
| *Ralstonia eutropha* | phaC$_{AC}$ | fatty acids | P(3HB:HHx) | 80 |
| *Ralstonia eutropha* | - | glucose | P(3HB:4HB) | 75 |
| *Ralstonia eutropha* | - | glucose + propionate | P(3HB:3HV) | 75 |
| *Escherichia coli* | phbCAB | glucose + 1,4-butanodiol | P(3HB:4HB) | 75 |
| *Escherichia coli* | phbCAB+vgb | glucose | PHB | 80 |
| *Pseudomonas putida* | - | fatty acids | PHA$_{MCL}$ | 60 |

This agrees with the fact that PHA accumulation is induced by nitrogen and/or phosphorus limitation, simply because a lack of these nutrients would be compensated by increasing the C, H and O contents of biomass (PHA is accumulated intracellularly).

This hypothesis implies that under phosphorus limitation, the proportion of fatty-acids in the phospholipids increases, decreasing that of phosphorus, which then can be routed to the RNA of ribosomes. So phosphorus limitation could stimulate up-regulation in the fatty-acids *de novo* synthesis in order to keep protein synthesis as fast as possible in a trade-off with the properties of the cell membrane.

If nitrogen is limiting but phosphorus is not, this effect would be counterproductive because it would not help protein synthesis and would still affect the properties of the cell membrane. Yet, when both nitrogen and phosphorus are limiting, the PHA contents observed (Taciro, 2008) suggest the metabolism is dominated by phosphorus limitation.

This could be due to the fact phosphorus is scarcer than nitrogen. For example, for a cell containing 12 %wt of nitrogen and 2 %wt of phosphorus with both of these nutrients on the verge of becoming limiting, a decrease of 1 g in the availability of phosphorus will have a much larger impact than the same decrease of 1 g in the availability of nitrogen.

# 4 MATERIALS & METHODS

The *N*-phenotypic model was developed during this very work, so in section 4.1 it is briefly described only as a method used to select the best points from the chemostat culture database to be later used in the *N*-GlobalFit automatic refining method (section 4.2) and also in the validation of the manually refined network from version 12 onwards. The theoretical properties themselves of the *N*-phenotypic model are discussed in section 5.1.

## 4.1 The *N*-phenotypic model

The "*N*-phenotypic model" was proposed to describe a common empirical pattern consistently reported in literature for *N*-datasets. It was developed together with the group European Research team in Algorithms and Biology, formaL and Experimental (ERABLE), in France. This model encodes the three metabolic Principles stated in section 3.2.1 and observed in the chemostat culture database of *Pseudomonas* sp. LFM046, namely:

I. First kinetic principle: the relationship between cell growth rate and substrate availability is best described by a saturation type of curve;

II. Continuity and smoothness: not one but multiple factors concomitantly impact the growth rate, and, moreover, the transition between growth regimes (set of impacting factors and how they impact growth) is continuous and smooth;

III. Multiple Objective (MO): organisms maximize their growth rate and also the consumption and storage of the most critical factors, in order to hinder competition.

*N* is the number of growth-limiting factors in a phenotypic *N*-dataset, such that this dataset allows the calculation of *N* specific inlet or outlet non-interchangeable mass flow rates to represent them. For example, a dataset with the volume of a batch bioreactor and concentrations of two non-interchangeable substrates has $N = 2$. A growth-limiting factor is not necessarily a substrate, it could be sunlight for an autotroph or an inhibitor. The outlet mass rate of biomass itself is not counted in *N*.

For this case-study of *Pseudomonas* sp. LFM046, $N = 3$. Definitions 1 to 4 introduce the notation used henceforth for limitation regimes. They are formalized in section 5.1.4.

**Definition 1**: a culture is in the {S1}-single-limitation regime when the non-interchangeable substrate S1 can be considered completely depleted and all $N$ yields $Y_{X/Si}$ , $i = 1 \ldots N$, can be considered constant ($i = 1 \ldots N$ includes substrate S1 itself).

**Definition 2**: the normalized uptake rate of substrate Si is $Z_i$ (Equation 7).

$$Z_i = \begin{cases} q_{Si}/q_{Sr} & , i \neq r \\ q_{Si}/c_{Sr} & , i = r \end{cases} \qquad \begin{array}{l} q_A\text{: specific uptake rate of A [g h}^{-1}\text{ (g cells)}^{-1}\text{]} \\ c_A\text{: specific supply rate of A [g h}^{-1}\text{ (g cells)}^{-1}\text{]} \\ \text{Sr: reference substrate} \end{array} \tag{7}$$

**Definition 3**: the $N$-dimensional vector space formed by all points $(Z_1, \ldots , Z_{N-1}, \mu)$ is named "$N$-Liebscher space" and said to use the "direct normalization" ($Z_i$). The $N$-dimensional vector space formed by all points $(1/Z_1, \ldots , 1/Z_{N-1}, \mu)$ is named "$N$-Egli space" and said to use the "inverse normalization" ($1/Z_i$).

**Definition 4**: an $N$-tuple multiple-limitation regime is "*N-lim*", "*{S1, ... , SN}-N-lim*" or "*S*$_a$*-N-lim*", being $S_a$ defined as the set of all substrates. Its complementary is "*0-lim*", "*{S*$_a$ *\ {S1, ... , SN}}-(N-k)-lim*" or "*∅-0-lim*".

An important remark is that for a mixture of two interchangeable substrates, like glucose and fructose (both are sources of carbon), Definition 1 can only be applied if both are represented by a single pseudo-substrate. This is the case of the chemostat culture database of *Pseudomonas* sp. LFM046.

### 4.1.1 A formulation for *N = 3*

The chemostat culture database of *Pseudomonas* sp. LFM046 is of class *3*-data: the measurements allow estimation of the specific growth rate ($\mu$) as well as the specific uptake rates of the 3 non-interchangeable substrates carbon ($q_C$), nitrogen ($q_N$) and phosphorus ($q_P$). Using Definitions 2 and 3, the *3*-Liebscher space is the set of points in Equation 8.

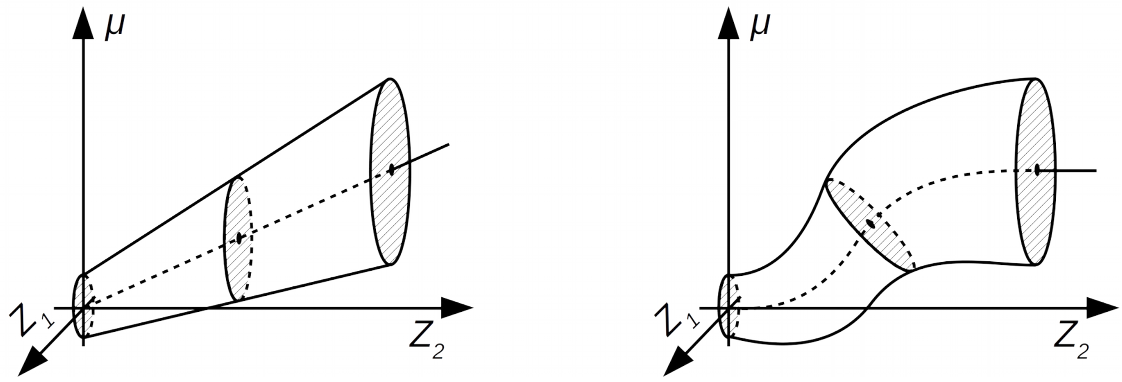
$$P \in \mathit{3}\text{-Liebscher} \Leftrightarrow P = (Z_1, Z_2, \mu) : \begin{cases} Z_1 = q_N/q_C \\ Z_2 = q_P/q_C \end{cases} \tag{8}$$

Then, applying Definition 4, the *3*-lim regime a subset of the *3*-Liebscher space. A convenient model is the region inside a 3D shell which is an elliptical cone whose apex is not a point but a small ellipse and whose axis is not a line but a sigmoid *1*-D curve, such that:

- the apex is at the origin of the 3-Liebscher space;
- the sigmoid axis is coplanar with the $\mu$ axis of 3-Liebscher space;
- the elliptic directrix is contained in a plane $H_p$ defined by a point $H_{pp}$ and a normal vector $H_{pn}$ in *3*-Liebscher space according to Statement 9.

$$\begin{gathered} H_{pp} \in \text{sigmoid axis} \wedge H_{pp} \text{ is centre of elliptic directrix} \\ H_{pn} \parallel \text{tangent vector of sigmoid axis at } H_{pp} \\ \text{the radii of the elliptic directrix increase with } \vec{Z} = (Z_1, Z_2) \end{gathered} \tag{9}$$

Thus, the generatrices of the 3D shell are not lines but *1*-D curves embedded in 3D space. Also, at an infinite distance from the $\mu$ axis in the first orthant (positive $Z_1$ and $Z_2$), $H_{pn}$ is orthogonal to $\mu$. Each elliptical directrix has one diameter aligned with the $\mu$ axis, so the other is contained in a plane parallel $Z_1 \times Z_2$. And finally, these diameters increase with the position in the *1*-D sigmoid axis according to a Monod model (Figure 6).



a) 3D linear cone in *3*-Liebscher space b) the proposed *3*-phenotypic model

*Figure 6: The shape of the proposed 3-phenotypic model, constructed by distorting a cone.*

The biological assumptions behind the shape depicted in Figure 6 are discussed in section 5.1.5. They make the shell symmetric around the sigmoid axis:

I. The ideal growth stoichiometry is constant (the sigmoid axis is coplanar with $\mu$);

II. The energy storage capacity is equal to the maximum maintenance energy;

III. The capacity to consume $S_i$ to save $S_j$ is equal to that of doing the opposite;

IV. The capacities in Assumptions II and III increase with the availability of the normalized factors ($||\vec{Z}||$) in the proportion of the ideal growth stoichiometry according to a Monod model (a saturation curve that starts almost linear).

Equation 10 defines the vector of 11 parameters for the *3*-phenotypic model. Equations 11 to 13 define the sigmoid 1-D path *PΣ* as a set of 3 parametric equations, one for each dimension in 3-Liebscher space). Equations 14 to 16 define the two unitary vectors of the directions of the diameters of the elliptical cross-section at every position in *PΣ*. Equations 17 and 18 define the size of these diameters and Equation 19 defines the shell inside which lies the triple-limitation region, that is, the 3D boundaries of the *3-lim* regime.

$$\vec{p} = (\mu_{max}, K_1, K_2, p_1, p_2, r_{x',0}, r_{x',max}, r_{x',s}, r_{y',0}, r_{y',max}, r_{y',s}) \tag{10}$$

$$\begin{aligned} P\Sigma = \quad & \{[P\Sigma_1(\alpha_1), P\Sigma_2(\alpha_1), P\Sigma_3(\alpha_1)] \in \mathbb{R}^3\} : \\ & \vec{\theta}_\Sigma = \tfrac{\vec{t}}{||\vec{t}||} : \vec{t} = \left(\tfrac{dP\Sigma_1}{d\alpha_1}, \tfrac{dP\Sigma_2}{d\alpha_1}, \tfrac{dP\Sigma_3}{d\alpha_1}\right) \end{aligned} \tag{11}$$

$$\{P\Sigma_1(\alpha_1) = K_1\alpha_1 \ ; \ P\Sigma_2(\alpha_1) = K_2\alpha_1\} \Rightarrow \vec{t_\Sigma} = \left(K_1, K_2, \frac{dP\Sigma_3}{d\alpha_1}\right) \tag{12}$$

$$P\Sigma_3(\alpha_1) = \frac{\mu_{max}}{1 + exp(-4p_1/\mu_{max}K_1(\alpha_1 - 1)) + exp(-4p_2/\mu_{max}K_2(\alpha_1 - 1))} \tag{13}$$

$$\vec{x'} = \frac{\vec{v_1}}{||\vec{v_1}||} : \vec{v_1} = (-K_1, K_2, 0) \tag{14}$$

$$h(w) = 2K_1\left(\tfrac{dP\Sigma_3(\alpha_1)}{d\alpha_1}\Big|_w\right)^{-1} \tag{15}$$

$$\vec{y'} = \frac{\vec{v_2}}{||\vec{v_2}||} : \vec{v_2} = (K_1, K_2, -h(\alpha_1)) \tag{16}$$

$$r_{x'}(\alpha_1) = r_{x',0} + r_{x',max}\ \alpha_1/(r_{x',s} + \alpha_1) \tag{17}$$

$$r_{y'}(\alpha_1) = r_{y',0} + r_{y',max}\ \alpha_1/(r_{y',s} + \alpha_1) \tag{18}$$

$$\vec{P_3}(\alpha_1, \alpha_2) = \vec{P\Sigma}(\alpha_1) + \vec{x'}\ r_{x'}(\alpha_1)\ cos(\alpha_2) + \vec{y'}(\alpha_1)\ r_{y'}(\alpha_1)\ sin(\alpha_2) \tag{19}$$

Therefore, if the 11 parameters of Equation 10 produce an accurate model of the *3-lim* regime of *Pseudomonas* sp. LFM046, then all points in *3*-Liebscher space which represent well cultures of this organism lie inside the 3D shell defined by Equation 19. Conversely, if a set of parameters produces a shell that circumscribes *X*% of the points calculated from experimental measurements, the other *(100-X)*% can be considered outliers. These 11 parameters were fit by a manual trial-and-error procedure visualizing the shell in a 3D plot.

## 4.2 The genome-scale metabolic network

Two methods to simultaneously build and validate the genome-scale metabolic network of *Pseudomonas* sp. LFM046 were used: the manual method and the automatic software *N*-GlobalFit, developed here. Both with the same kinds of inputs, namely:

I. a draft genome-scale generated automatically by the webservice KBase, from the genome assembly generated with the software MeDuSa (Bosi et al, 2015). In KBase, biomass production was ensured without restricting any exchange reaction;

II. a list of potential reactions to be added to the draft network. Each reaction has a unique ID, a name, an equation with also unique metabolite ID's and a penalty score. This score is to prioritize which reactions to add;

III. a list of *3*-phenotypic points. Each point has 8 fluxes calculated from the experimental *3*-data: $q_{glu}, q_{fru}, q_{NH4}, q_{PO4}, q_{O2}, q_{PHA}, q_{CO2}$ and $\mu$. The first 5 are substrates, of which the first two are used to calculate $q_C$ (carbon uptake rate), $q_{NH4} = q_N$ and $q_{PO4} = q_P$. $N = 3$ because this data describes the *{C, N, P}-3-lim* regime.

The manual method consists in finding and solving one inconsistency at a time. The experimental *3*-data were generated in chemostat cultures with defined media, so it is known that *Pseudomonas* sp. LFM046 grows and how fast it grows in these media. Thus, most of the inconsistencies can be found simply by constraining some or all of the 8 measured fluxes and running an FBA simulation.

For example, if the FBA indicates that the strain cannot grow without consuming external phenylalanine and that was not in the culture medium, then the production pathways of phenylalanine are analysed backwards, one reaction at a time, until a gap is found. Then, the list of potential reactions is consulted to find reactions that could fill that gap. This is the same idea of the gapfilling algorithms discussed in section 3.3.1.

The manual method was used as a control for *N*-GlobalFit, taking into account the time to carry-out, the complexity of the human intervention and the predictive quality of the resulting metabolic network. The latter is to compare the proportion of right and wrong predictions (inside and outside the 3D shell of the *3*-phenotypic model).

### 4.2.1 *N*-GlobalFit

The software *N*-GlobalFit is written in the R programming language and is an adaptation of the GlobalFit library. No changes were needed in the core code, because the adaptations could be implemented by conveniently preparing the input data. The GlobalFit algorithm and its input data types are described in section 3.3.1, and the formal mathematical definitions were published in 2016 (Hartleb et al, 2016). The adaptations were:

- As the “on” list of growth experimental points, set a list of points in the lower surface of the 3D shell (equivalent to the dashed curve in Figure 16b), with the value of $\mu$ in place of the growth viability threshold $T_g$;
- As the “off” list of growth experimental points, set a list of points in the upper surface of the 3D shell (equivalent to the solid curve in Figure 16b), with the value of $\mu$ in place of the non-growth viability threshold $T_h$;
- Extra linear restrictions like ratios between fluxes were set as artificial reactions and the “knock-out” list of each “on”/”off” point was used to disable them individually.

### 4.2.2 The database of potential reactions

In the publication of GlobalFit, the additional reactions are the BiGG database minus the reactions from GPR’s of dissimilar genes (BLAST *e*-value > $10^{-13}$) and then minus the blocked reactions in the supernetwork (Hartleb et al, 2016). For *Pseudomonas* sp. LFM046, there are fewer similar organisms and not all of them have well-curated published models from which reliable GPR’s can be extracted.

The two best models in the trade-off between genetic similarity and GPR reliability are iPAE1146 (Bartell et al, 2017) and PpuQY1140 (Yuan et al, 2017). So they were used to build the list of potential reactions. However, the first uses the metabolite ID naming system from modelSEED whereas the second uses its own system. Thus, both were first converted to the METANETX namespace version 3.1 (https://www.metanetx.org/):

- iPAE1146: 1284 metabolites, 1241 converted directly by ID (97%);

- PpuQY1140: 1104 metabolites, 1078 converted directly by name (98%).

In total, 69 metabolites required a deeper verification of the reactions they participate in and/or their structure to find their ID's in METANETX namespace. Then the two lists of reactions and their GPR's were merged into a single supernetwork. Each reaction was associated to a score calculated from the results of SonicParanoid (Cosentino et al, 2019):

- a list of gene orthologs between strains LFM046, *P. aeruginosa* PAO1 and *P. putida* KT2440 was generated from the protein sequences of these 3 organisms (FASTA files) by SonicParanoid;
- the score for each reaction in the supernetwork is $0.5 + 0.5 \times n_{orts} / n_{GPR}$, where $n_{GPR}$ is the number of genes in the GPR and $n_{orts}$ is the number of orthologs of these genes found in the genome of the strain LFM046. For GPR's with more than 1 set of genes, the highest score was selected. For example:
  - GPR: (G1 and G2 and G3) or (G1 and G4)
  - if there are orthologs for G1 and G3, the score will be $0.5 + 0.5 \times 2/3$

Then, duplicate reactions were removed by calculating a cosine similarity matrix from the supernetwork stoichiometric matrix (using the cosSparse function of the qlcMatrix package, in the R language) and verifying those reaction pairs with a similarity higher than 0.95 or lower than -0.95. Using the same method, 679 reactions present in the KBase draft (converted to METANETX) were also removed from the supernetwork, being 246 of them not exact matches.

Exchange reactions were removed from the supernetwork because the culture media of all experiments were defined. Reversible reactions were replaced by pairs of irreversible reactions. The final list of potential reactions has 1758 reactions (version 12b).

### 4.2.3 The list of phenotypic points

A large database of 44 chemostat experiments of *Pseudomonas* sp. LFM046 with an average duration of 100 h and a total of 258 observations (samples) in steady-state was the source of the phenotypic points. The average number of observations is 5.9 per steady-state and 0.39 per steady-state-hour. Most of these experiments were carried-out before 2008 (Taciro, 2008), so the first step was to review the values and original logs, including to re-evaluate whether an observation would be considered in steady-state or not. Each observation is a set of 22 measured variables, as described in Table 4.1.

*Table 4.1: Measured variables in the phenotypic database*

| ID | Name | Units |
|---|---|---|
| $t$ | Time | h |
| $V$ | Volume | mL |
| $F_{G,in}$ | Inlet gas flow rate | mL min$^{-1}$ |
| $F_{L,in}$ | Inlet liquid culture medium flow rate | mL min$^{-1}$ |
| $F_{pH}$ | Inlet liquid pH control flow rate (alkali and acid solutions) | mL min$^{-1}$ |
| $Ni_{in}$ | Inlet concentration of equivalent inorganic nitrogen | mg L$^{-1}$ |
| $Pi_{in}$ | Inlet concentration of equivalent inorganic phosphorus | mg L$^{-1}$ |
| $G_{in}$ | Inlet concentration of glucose | g L$^{-1}$ |
| $F_{in}$ | Inlet concentration of fructose | g L$^{-1}$ |
| $X_T$ | Outlet concentration of total biomass (cells + PHA) | g L$^{-1}$ |
| *C04* | 3-hydroxybutyrate content in total biomoass | %wt in $X_T$ |
| *C06* | 3-hydroxyhexanoate content in total biomass | %wt in $X_T$ |
| *C08* | 3-hydroxyoctanoate content in total biomass | %wt in $X_T$ |
| *C10* | 3-hydroxydecanoate content in total biomass | %wt in $X_T$ |
| *C12d5* | 3-hydroxy-Δ5-dodecenoate content in total biomass | %wt in $X_T$ |
| *C12* | 3-hydroxydodecanoate content in total biomass | %wt in $X_T$ |
| $Ni_{out}$ | Outlet concentration of equivalent inorganic nitrogen | mg L$^{-1}$ |
| $Pi_{out}$ | Outlet concentration of equivalent inorganic phosphorus | mg L$^{-1}$ |
| $G_{out}$ | Outlet concentration of glucose | g L$^{-1}$ |
| $F_{out}$ | Outlet concentration of fructose | g L$^{-1}$ |
| $y_{O2}$ | $O_2$ content in the offgas stream | %vol |
| $y_{CO2}$ | $CO_2$ content in the offgas stream | %vol |

Traditionally, the uncertainty of a variable calculated from experiments is estimated by uncertainty propagation theory taking into account the averages and the uncertainties of the measured variables. The latter is usually a function of the standard deviation calculated per variable over the set of all observations of it in the same steady-state (Taciro, 2008).

An alternative multivariate approach was adopted. Each observation was considered as a 21-D vector (all variables minus time, because the data is in steady-state) and for each coordinate, 100 estimations were calculated as in Equation 20, where $x$ is a coordinate (a variable), $\gamma$ is a random number between -0.5 and 0.5 and $\delta_x$ is the uncertainty of the variable $x$ estimated from equipment specifications and methodological tests.

$$x_{estimated} = x_{measured} + \gamma \cdot \delta_x \tag{20}$$

Then, for each one of the 25800 simulated 21-D observations, the 8 rates $q_{glu}$, $q_{fru}$, $q_{NH4}$, $q_{PO4}$, $q_{O2}$, $q_{PHA}$, $q_{CO2}$ and $\mu$ were calculated, as well as two carbon mass balance error estimators. For each one of these 10 calculated variables, the percentiles 97.5, 50.0 and 2.50 were selected as the maximum, average and minimum values of each steady-state. This is a kind of Monte-Carlo simulation, carried-out in spreadsheet software (LibreOffice Calc).

The carbon mass balance error estimators were $\epsilon_E$ and $\epsilon_L$. They are defined in Equations 21 and 22, where each coordinate $q_i$ is the specific consumption or production rate of the $i$-th metabolite, the suffix $S$ denotes the subset of substrates and $\vec{y}_C$ is the vector of carbon mass fractions in each metabolite (e.g. zero for $NH_4^+$ and ca. 0.4 for glucose).

$$\epsilon_E = \frac{\langle \vec{q}_{in}, \vec{y}_C \rangle - \langle \vec{q}_{out}, \vec{y}_C \rangle}{\langle \vec{q}_{S,in}, \vec{y}_{C,S} \rangle} \tag{21}$$

$$\epsilon_L = \frac{\langle \vec{q}, \vec{y}_C \rangle}{\langle \vec{q}_S, \vec{y}_{C,S} \rangle} \tag{22}$$

$$q_i = q_{in} - q_{out} \tag{23}$$

Thus, $\epsilon_E$ is calculated for the bioreactor and is relative to its total carbon supply whereas $\epsilon_L$ is calculated for the cell population and is relative to its total carbon consumption rate. The numerators of both are equal due to Equation 23. Equation 24 is the ratio $\epsilon_E / \epsilon_L$ for two carbon sources S1 and S2.

$$\epsilon_{EL} = \frac{q_{S1,in} y_{C,S1} + q_{S2,in} y_{C,S2} - q_{S1,out} y_{C,S1} - q_{S2,out} y_{C,S2}}{q_{S1,in} y_{C,S1} + q_{S2,in} y_{C,S2}} \tag{24}$$

All values are positive (mass fractions and rates), so $\epsilon_E / \epsilon_L \leq 1$. For this reason, $\epsilon_L$ was adopted as the multivariate uncertainty estimator. When $\epsilon_E / \epsilon_L = 1$, all carbon supplied to the bioreactor is consumed, meaning that carbon is either a limiting substrate or is being supplied at the exact rate demanded by the population. Otherwise it is proportional to the normalized concentration *Zr* for the case $N = 1$ of the phenotypic model.

For each one of the 258 points, $q_C$, $q_N$ and $q_P$ were calculated to plot the *3*-Liebscher space of points $(q_N/q_C,\ q_P/q_C,\ \mu)$. The 3D shell from Equation 19 was fit visually to these points by trial-and-error on the 11 parameters, using the Python language (Matplotlib and Jupyter Notebook). The criterion was to maximize the number of the points inside the shell but still minimize the distances between it and the most external points inside of it (tightly fit). This can be formalized and solved by software, but that was not necessary because the outliers (literally) were few and very clear.

Then, 21 points near the shell were selected to represent it, 11 from the upper side of and 10 from the lower side. The average $\epsilon_L$ for these 21 points was 17%, so a margin of 20% was adopted for the values of the $T_g$ and $T_h$ thresholds (for the *i*-th point among 10 lower-side, $T_{g,i} = 0.8\ \mu_{min,i}$ and for the *i*-th point among the 11 upper-side, $T_{h,i} = 1.2\ \mu_{max,i}$). The shell could not be used directly because there is no simple way to determine which set of glucose and fructose uptake rates corresponds to a value of $q_C$. This is the list of phenotypic points to be input in *N*-GlobalFit.

For the manual method up to version 12 of the metabolic network, the list only had 5 points, of which the only common ones with the 21 from the 3D shell are 2 points from the upper side. The 5 points were chosen not for being the most external ones but for being from distinct limitation regimes, and were calculated not from single 21-D observations for each but as 21-D averages representative of the whole steady-state. The logic was to mimic how the manual process is usually carried-out: few points, without the theory of the *N*-phenotypic model and with single-variate averaged values.

## 4.3 Hypotheses and simulations

There is a large database of steady-state data for *Pseudomonas* sp. LFM046, some data in transient-state (the continuous cultures before stabilization) and some data of recombinants of this strain. There are no estimates of specific kinetic parameters and not enough transient-state data for dynamic simulations (e.g. stochastic methods or dynamic Flux Balance Analysis – dFBA), so they were ruled out.

According to section 3.4.1, structural analysis like Elementary Flux Mode Analysis (EFMA) was evaluated but found to be impractical and less adequate than constraint-based modelling for the case-study of *Pseudomonas* sp. LFM046. With that, Flux Balance Analysis (FBA) was the adopted technique, as it was the best alternative in the trade-off between simplicity and potential insights. Aside from the manual network refining method (section 4.2), FBA was also used to:

- test the primary and secondary hypotheses (sections 4.3.1 and 4.3.2);
- evaluate the maximum theoretical PHA content in function of its composition;
- decide which should be the next experiment and why.

OBS: Flux Variabiliry Analysis (FVA) is a set of FBA simulations. It was also used, but only to verify the results of the tests using FBA and find mistakes in the formulations of the simulations, so the FVA results were omitted from section 4. The commands used to perform the simulations in the webservice F.A.M.E. are transcribed in Appendix B.

### 4.3.1 Primary hypothesis

The primary hypothesis is: there is an Energy-Dissipating Cycle (EDC) of fatty-acids *de novo* synthesis and β-oxidation, caused by the imbalance between their high speed and the low speed of the recombinant $PHA_{SCL}$ polymerase. The equivalent turnover of this cycle is much lower or even zero in the wildtype strain LFM046 because its native $PHA_{MCL}$ polymerase is fast enough to consume the $HA_{MCL}$ monomers (mostly $C_{10}$).

The compatibility of the genome-scale metabolic network of *Pseudomonas* sp. LFM046 with this hypothesis was tested by the following set of FBA simulations:

I. the flux in a reaction that produces $C_{10}$ from $C_8$ (R_rxn05343_c0) was maximized and from that was subtracted the resulting flux of the production of Acetyl-CoA in the end of β-oxidation (-R_rxn00178_c0, that is, in the opposite direction);

II. the flux of the artificial reaction A_BIOTOT01_u was constrained to distinct values (this is the independent variable). This reaction sets the proportion between PHA and biomass and is from an experimental steady-state (ca. 55 %wt of PHA);

III. the fluxes of the exchange reactions were constrained to ranges estimated from the experimental point encoded in A_BIOTOT01_u (the point #20108), which was limited in phosphorus but not on nitrogen nor carbon (energy). These constraints are:

- EX_cpd00027_e0: glucose, from -20 to 0,

- EX_cpd00082_e0: fructose, from -20 to 0,

- EX_cpd00007_e0: $O_2$, from -50 to 0,

- EX_cpd00013_e0: $NH_4^+$, from -15 to 0,

- EX_cpd00009_e0: $PO_4^{3-}$, from -1 to 0,

- EX_cpd00011_e0: $CO_2$, from 50 to 100 (the only product, the others are substrates);

IV. the stoichiometry of the artificial reaction R_A_PHA_SELECTOR was set to two cases. One for a PHA with 95 %mol of $C_4$ and the other for the PHA composition of the wildtype strain LFM046, which has no $C_4$ and has over 60 %mol of $C_{10}$.

Care must be taken with the maximization in item I: it is not the same as maximizing $v_{ff} = v_{R\_rxn05343\_c0} - v_{R\_rxn00178\_c0}$, where $v_{ff}$ is the equivalent turnover of the fatty-acids futile cycle. This could simply be compensated by any Energy-Generating Cycle (EGC) present in the network. Maximizing only $v_{R\_rxn05343\_c0}$ *does not guarantee* robustness against EGC’s, but if the results indicate a linear relationship between $v_{ff}$ and $v_{A\_BIOTOT01\_u}$, then it is proved that the network *is able to predict* this fatty-acids futile cycle *even without any EGC*.

### 4.3.2 Secondary hypothesis

The secondary hypothesis is that the turnover in the fatty-acids *de novo* synthesis increases as a response to carbon excess, and, additionally, under phosphorus limitation, it increases relatively more as if it were up-regulated or if its competing pathways were all down-regulated. This could explain the following observations:

- the wildtype strain *Pseudomonas* sp. LFM046 with its PHA rich in $C_{10}$ reaches very similar PHA contents in the regimes *P-1-lim* and *{N, P}-2-lim* but significantly lower contents in *N-1-lim*;
- a recombinant of this strain with the same central metabolism but with a PHA poor in $C_{10}$ reaches reaches a much higher content in *N-1-lim* than in *P-1-lim*, and both much lower than those of the wildtype;
- the wildtype strain consumes almost all glucose before starting to consume fructose when growing fast and completely inverts this diauxic growth when phosphorus becomes limiting and growth is replaced by PHA accumulation (fed-batch culture).

The first part of the hypothesis was modelled by having more than one reaction forming biomass, with slightly different compositions, in order to match the nitrogen and phosphorus mass balances of the chemostat cultures in distinct limitation regimes. Since the total biomass composition must add up to 100%, a lack of N or P must be compensated by other elements. Trace elements and S would have to increase too much for that (more than double), so they are ruled out, leaving C, H and O. Thus:

- *P-1-lim*: the lack of P could be compensated by C, H, O and/or N;
- *N-1-lim*: the lack of N could be compensated by C, H, O and/or P;
- *{N,P}-2-lim*: the lack of both N and P could only be compensated by C, H and/or O.

The biomass macro-components richest in C, H and O are carbohydrates, fats and their precursors/derivatives (e.g. PHA is a derivative of precursors of fats). So the alternative biomass reactions necessarily increase the relative anabolic turnovers of these kinds of precursors. That is the very essence of the first part of the secondary hypothesis: speeding up the fatty-acids *de novo* synthesis as well as pathways which compete with it for carbon.

The rationale behind the second part is to prioritize the fatty-acids *de novo* synthesis among all these possible pathways. This would explain why this strain accumulates more PHA when phosphorus is limiting regardless of whether nitrogen is limiting too, what agrees with the observations in section 3.4.1.

The secondary hypothesis was evaluated in the genome-scale metabolic network of *Pseudomonas* sp. LFM046 with FBA simulations, by constraining the BIOTOT flux (biomass plus a fixed proportion of PHA) and minimizing the total carbon uptake flux with different glucose/fructose ratios. The evaluated response was the flux of the membrane-bound NAD/NADP transhydrogenase, which consumes proton-motive force to oxidise NADH and reduce $NADP^+$. The rationale is:

- Under nitrogen and/or phosphorus limitation, the biomass composition forces an increase in the turnover of any pathway that produces {C, H, O}-rich metabolites;
- There are several pathways that can supply this extra demand. One is the fatty-acids *de novo* synthesis. But slow growth causes $C_{10}$ to accumulate, so the β-oxidation is activated as well, forming a futile cycle: $NADPH + NAD^+ \rightarrow NADP^+ + NADH$;
- Glucose and fructose are metabolized by distinct pathways, such that the latter produces less NADPH per mole of carbon than the former. Thus, if their fluxes into the network are both minimized and their ratio is constrained to different values, the more fructose, the less NADPH for the fatty-acids futile cycle;
- The membrane-bound transhydrogenase spends energy to do the opposite of the fatty-acids futile cycle. Thus, if the fatty-acids *de novo* synthesis is prioritized, the flux distributions resulting from FBA minimizing the total carbon uptake rate should show that the flux of the reaction of the membrane-bound transhydrogenase increases with the proportion of fructose consumed by the network, without any extra restrictions forcing this. Additionally, phosphorus should be more limiting than nitrogen.

# 5 RESULTS & DISCUSSION

The first general result is a criterion to objectively assess the reliability of the data generated in a bioreactor system, based not on mathematical concepts like standard deviation and Principal Component Analysis nor on broad principles like mass balance but on the *N*-phenotypic model, which is founded on specific biological theory. The *N*-phenotypic model is a relationship between the specific uptake rates of *N* non-interchangeable substrates and the specific growth rate, forming an *N*-dimensional region of all possible phenotypes. Figure 7 shows that for *N* = *3*, the double-limitation regimes (*2-lim*) are almost planar subregions.

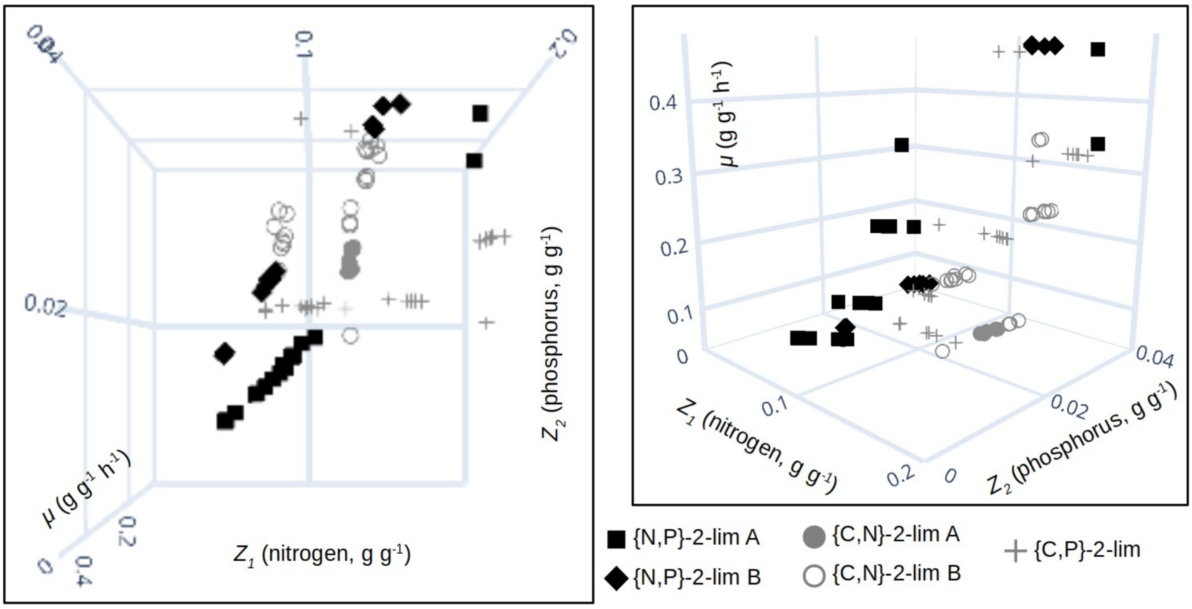


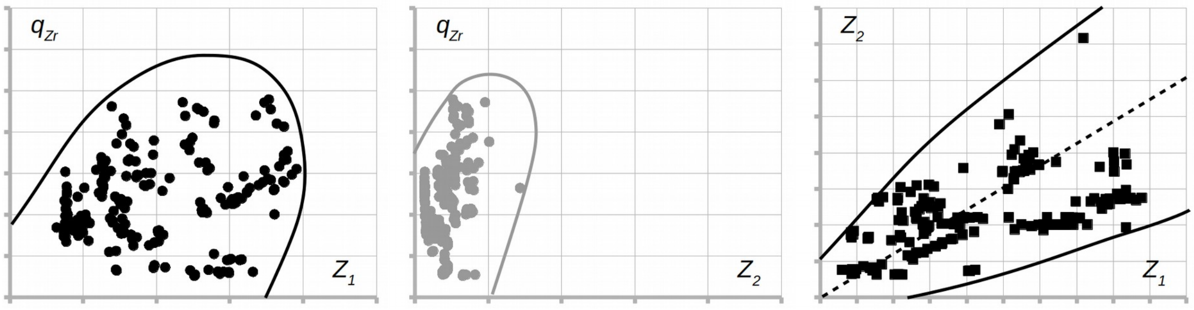
*Figure 7: Points of double-limitation regimes of Pseudomonas sp. LFM046 in 3-Liebscher space.*

Moreover, the direction of the normal vector of a plane that approximates a *2*-lim regime determines how relatively limiting each nutrient is. For example, the difference between the sets A and B of *{N, P}-2-lim* is that cultures in A were fed with more phosphorus and less nitrogen for the same amount of carbon, so the set A is closer to the nitrogen axis ($Z_1$). Analogously, the *{C, N}-2-lim* sets are almost parallel to the phosphorus axis ($Z_2$) because in these cultures nitrogen was always in excess.

Each of these *3*-Liebscher points was calculated from experimental *3*-data containing random and systematic errors, being the variables $Z_1$ and $Z_2$ non-linear on the measured variables (they are ratios of uptake rates). The probability of a set of random points of these measured variables forming a plane in *3*-Liebscher space is very low, so the overall quality of the experimental data can be estimated as a function of an error estimator of the multivariate linear regression of double-limitation points in *3*-Liebscher space.

Certainly, the higher *N*, the more information is encoded in the *N*-Liebscher space. Thus, the less likely it is that random variables generate points distributed in a non-uniform way in this vectorial space. This is true for the planes of double-limitation shown in Figure 7 and also for the shell of the *N*-lim itself. Each cross-section of the *N*-phenotypic model is a *N*-ellipsoid which can be transformed into an *N*-ball (sphere of unitary radius) with convenient scaling. But it is known that *V(N*-ball*)/V(N*-cube*)*, where *V* is *N*-volume, decreases asymptotically to zero as *N* grows (Parks, 2013). This allows to develop an automated method of bias evaluation and unbiased data reconciliation for any *N*-dataset.

The second general result is that if energy is indeed a factor much more limiting than the chemical nature of any non-interchangeable substrate, then the projection of the *N*-lim shell onto the subspace of normalized uptake rates ($Z_i$, all dimensions except $\mu$) forms a *(N-1)*-region unbounded in 1 path (if it is linear, it can be described as "one direction"). This is exemplified for *N =3* in Figure 8.



a) Nitrogen and carbon b) Phosphorus and carbon c) Nitrogen and phosphorus

*Figure 8: Pareto-regions of the trade-off between maximizing the growth rate and the substrate utilization for pairs of non-interchangeable substrates consumed by Pseudomonas sp. LFM046:* $q_{Zr}$ *is the carbon uptake rate* ($g\ g^{-1}\ h^{-1}$), $Z_1$ *is the normalized uptake rate of nitrogen and* $Z_2$ *is of phosphorus* (($g\ g^{-1}\ h^{-1}$) ($g\ g^{-1}\ h^{-1}$)$^{-1}$).

Carbon is the only energy source among the *3* nutrients and the regions of multi-objective optimality involving it (Pareto-region) are bounded (Figure 8a and b), but the region not involving it (Figure 8c) has at least one direction for which the consumption rate of all substrates can increase indefinitely, at least within practical ranges. In that direction, there is no practical trade-off between consuming nitrogen and phosphorus, because that is the ideal growth stoichiometry in which growth is only limited by energy. In other words, the dashed line is an approximation of a projection of the *0-lim* regime.

Figure 8c is an evidence that, at least for *Pseudomonas* sp. LFM046, the ideal growth stoichiometry can be considered constant for all its possible phenotypes. However, this stoichiometry is only equivalent to the biomass composition if the changes in the proportion and composition of the excreted products along the path of ideal stoichiometry are either negligible or symmetric around it (that is, if they compensate each other).

The third general result is that for any unknown chemoorganotroph, the substrates which probably generate the most informative *N*-phenotypic model with the smaller amount of measurements are all those containing carbon, nitrogen and a third element except hydrogen and oxygen, since:

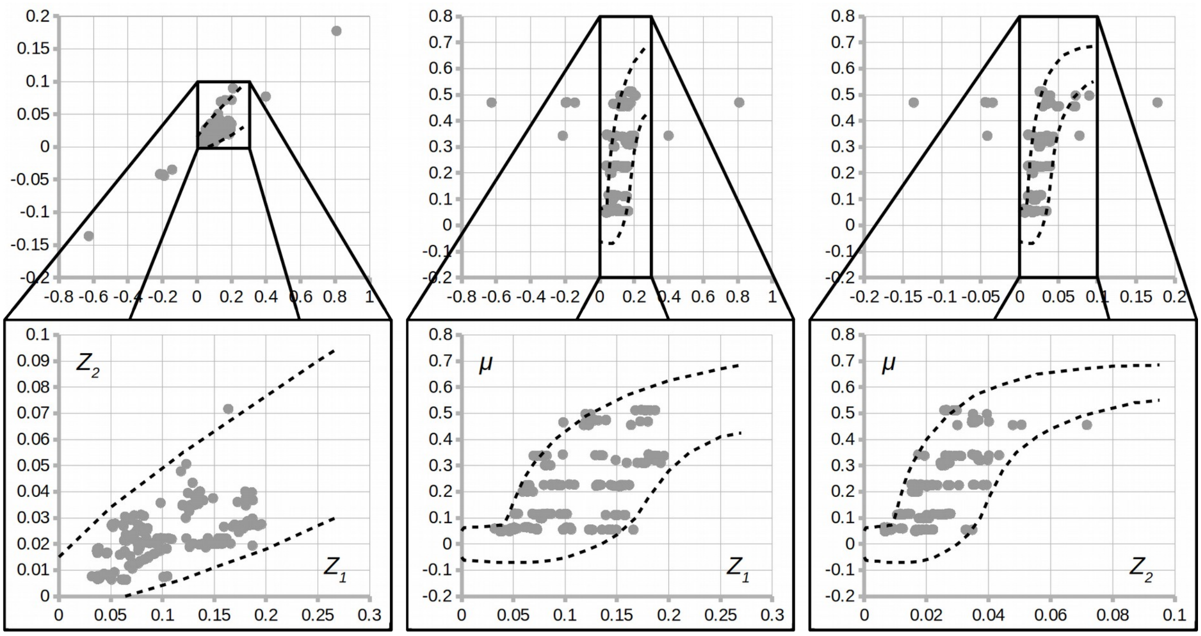
- $q_H$ and $q_O$ are linearly dependent on $q_C$, because protons related to the proton-motive force and oxygen related to respiration are excluded from them (those parts of the uptake rates are correlated to the energy source);
- without the third nutrient the model is 2D and the first general result cannot be used;
- a fourth nutrient will necessarily be a minor one with little impact in the path of ideal stoichiometry and in the flexibility around it. The flexibility is to consume more of one abundant substrate relative to the ideal stoichiometry in order to consume less of another scarce substrate relative to the same ideal stoichiometry. In Figure 8c this is any line orthogonal to the dashed line within the area between the solid curves.

The following subsections present results that are specific to the case-study of *Pseudomonas* sp. LFM046 with the same order as section 4: the *N*-phenotypic model, the genome-scale metabolic network and the simulations. The latter comprises the evaluation of the primary and secondary hypotheses as well as the resulting propositions for optimization of the production of PHA with control of the monomeric composition. A side note on modelling the maintenance energy in the metabolic network is also provided.

## 5.1 The *N*-phenotypic model

Out of 258 points of *3*-data of *Pseudomonas* sp. LFM046, 233 are inside, on or near the outer 3D shell which was fit to this dataset by a simple manual trial and error procedure with visual inspection in the 3D plot. Thus, less than 10% of the points were discarded (they were very far from the shell and/or had negative *Z* values due to gross measurement errors). Figure 9 presents the 2D orthogonal projections of the 258 initial points, the 233 filtered points and the 3D shell with its 11 parameters, all in *3*-Liebscher space.

a) $Z_1 \times Z_2$, initial dataset b) $Z_1 \times \mu$, initial dataset c) $Z_2 \times \mu$, initial dataset



d) $Z_1 \times Z_2$, filtered dataset e) $Z_1 \times \mu$, filtered dataset f) $Z_2 \times \mu$, filtered dataset

$Z_1$ and $Z_2$: nitrogen and phosphorus normalized uptake rates [(g g$^{-1}$ h$^{-1}$) (g g$^{-1}$ h$^{-1}$)$^{-1}$]
$\mu$: specific growth rate (h$^{-1}$)

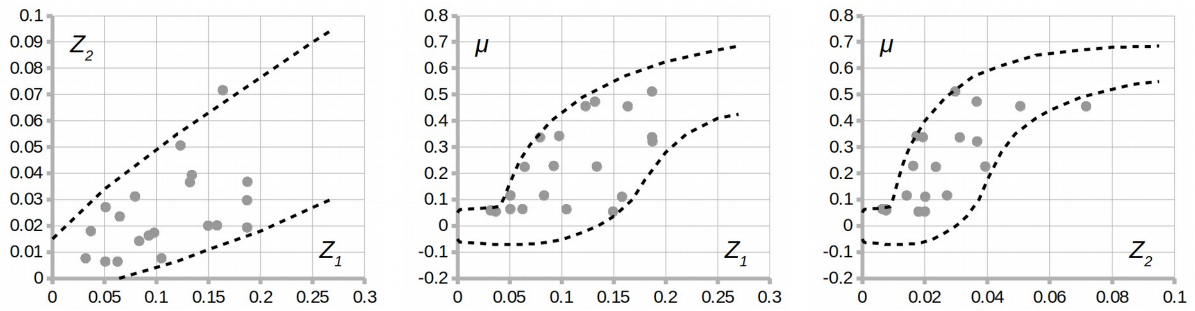
Vector of parameters of the fitted 3D shell, as in Equation 10 (dashed curves, Equation 19)

| $\vec{p}$ = | ( $\mu_{max}$ | $K_1$ | $K_2$ | $p_1$ | $p_2$ | $r_{x',0}$ | $r_{x',max}$ | $r_{x',s}$ | $r_{y'0}$ | $r_{y',max}$ | $r_{y',s}$ ) |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | 0.64 | 3/25 | 2/75 | 12 | 10 | 0.02 | 0.1 | 2 | 0.06 | 0.06 | 2 |

*Figure 9: A 3-phenotypic model fitted to 233 points of the {C, N, P}-3-lim regime of Pseudomonas sp. LFM046. The 25 discarded outliers are the points far from the region between the dashed curves (a, b and c).*

The average magnitude of the carbon mass balance error $\epsilon_L$ in the initial dataset was 25.5%. In the filtered set of 233 points, it dropped to 16.1%. In the 21 points selected to represent the 3D shell, it was 17.0%. These 21 points were selected visually in 3D, trying to maximize the distance between them while minimizing the sum of orthogonal distances from them to the surface of the shell.

Figure 10 presents the set of 21 points selected to represent the 3D shell. Each value of $q_C$ in the denominators of $Z_1$ and $Z_2$ is a sum of two values from glucose and fructose, $q_{C,glu}$ and $q_{C,fru}$, and there is no simple way to estimate them from a value of $q_C$. Thus, selecting representative points that already contain that information avoids considering a single artificial carbon source in the next steps of validation and refining of the metabolic network. These points also contain PHA and $CO_2$ production rates as well as $O_2$ uptake rate (section 4.2), which are not considered in the current version of the *N*-phenotypic model.



a) $Z_1 \times Z_2$ b) $Z_1 \times \mu$ c) $Z_2 \times \mu$

$Z_1$ and $Z_2$: nitrogen and phosphorus normalized uptake rates [(g $g^{-1}$ $h^{-1}$) (g $g^{-1}$ $h^{-1}$)$^{-1}$]
$\mu$: specific growth rate ($h^{-1}$)

*Figure 10: The 21 points selected to represent the 3-phenotypic model of Pseudomonas sp. LFM046.*

Comparing Figure 10c with Figure 9f it is visible that the 3-phenotypic model was fitted to points near *($Z_2$, $\mu$) = (0.04, 0)* which were not selected to represent the 3D shell. These points had a high $\epsilon_L$ error. Without them, the projection of the lower part of the shell onto the $Z_2 \times \mu$ plane would be different and could even place the point near *(0.07, 0.5)* inside the shell. However, that point also has a high $\epsilon_L$ error and was not discarded only because there was no other point near that part of the shell.

Considering the average $\epsilon_L$ of 17%, it was adopted the tolerance of 20% to compare the 21 representative points with predictions of the network, at least for the first comparisons. In Figures 10b and c, it is also clear that if the lower part of the shell is biased, it is such that it relaxes the problem of fitting the predictions of the metabolic network inside the shell.

### 5.1.1 The three metabolic Principles

The case $N = 2$ is analysed to assess how the three metabolic Principles of section 3.2.1 are encoded in the $N$-phenotypic model. Principle I is the sigmoid axis, the "saturation type of curve". Adding Principle II, this curve must be continuous and differentiable. To plug-in Principle III, first the concept of Pareto frontier (Gorban et al, 2011) is shown in Figure 11.

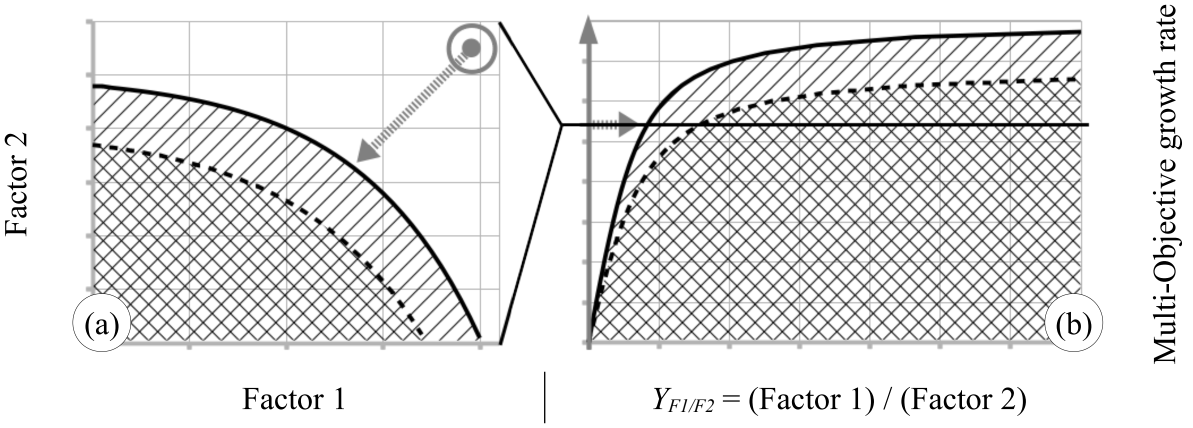

Factor 2
(a)
Factor 1
(b)
Multi-Objective growth rate
$Y_{F1/F2}$ = (Factor 1) / (Factor 2)

*Figure 11: Projection of two Pareto frontiers for the availabilities of two growth-limiting factors (a, adapted from Gorban et al, 2011) and its corresponding plot of the multi-objective specific growth rate in function of the ratio between the availabilities two factors (b).*

In Figure 11a (left) it is depicted the concept of a Pareto frontier. Each one of the dashed and solid black curves partitions the 2D space in two: inside the hatched area both factors can change without decreasing the multi-objective function, outside they cannot. This is the behaviour of the "generalized Liebig's systems", with a convex "area of better conditions" (Figure 2b of Gorban et al, 2011).

In Figure 11b (right), the horizontal axis encodes the Pareto frontiers from Figure 11a and the vertical axis is simply a linear function representing the specific growth rate. It is no coincidence that this produces a saturation curve. In Principle I, it is tacitly assumed that the substrate availability is normalized, that is, all points correspond to populations of the same size or, equivalently, to the same availability of a growth-limiting substrate of reference.

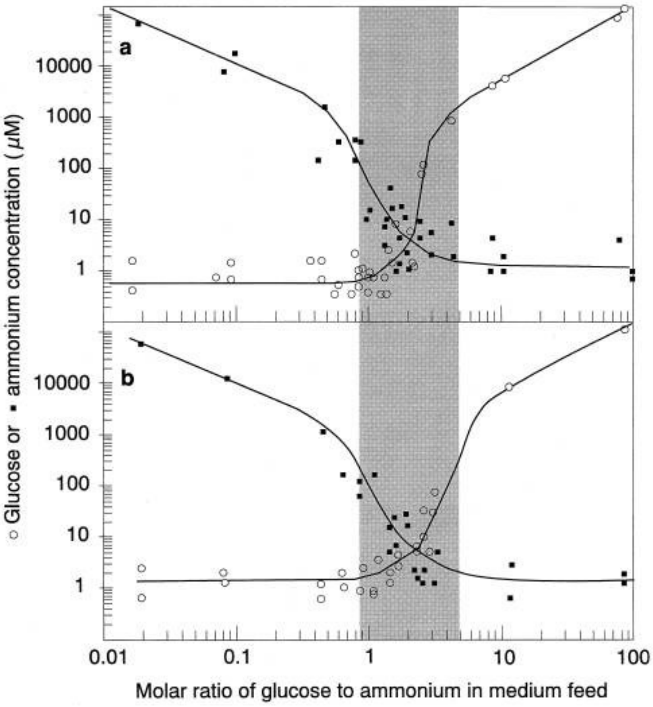
Still in Figure 11b, the curves represent the maximum possible specific growth rate, $\mu$, given that only the most critical factor of the two is growth-limiting (dashed) or that only the least critical is (solid). In the area between the curves, both factors are concomitantly limiting, so the growth rate is not as high as in the solid curve nor as low as in the dashed one.

This introduces the concept of continuous factor limitation: a factor is not simply "limiting" or "non-limiting", but instead it is more or less limiting than another factor. This is Principle II applied to the set of factors needed for growth. Traversing the horizontal path from left to right in Figure 11b, three classes of limitation regimes are characterized:

i. region without hatching, low $Y_{F1/F2}$: factor 1 (F1) is scarce relative to factor 2 (F2). So, F1 is more "limiting" than F2, or F2 is "saturated" in relation to F1. Thus, the biomass yields $Y_{X/F1}$ and $Y_{X/F2}$ are both constant along the path, analogously to a single chemical reaction in which F1 is the stoichiometrically limiting reactant;

ii. region with single hatching + the two curves, medium $Y_{F1/F2}$: both factors significantly limit the growth rate. The biomass yields $Y_{X/F1}$ and $Y_{X/F2}$ respectively decrease and increase along the path (Figure 12 with nitrogen = F1 and carbon = F2);

iii. region with double hatching, high $Y_{F1/F2}$: analogous to (i) with F1 and F2 switched.



The plot on the left is Figure 7 from Kovárová-Kovar, 1998. It shows the outlet concentrations of glucose and ammonium in function of the glucose/ammonium concentration ratio in the inlet of chemostat cultures of two dilution rates (a and b, 0.2 and 0.4 $h^{-1}$) of the same bacterial strain.

All biomass yields are constant in the regions where the glucose and nitrogen concentrations both vary linearly. These are the regions of C-single-limitation (left) and N-single-limitation (right). The grey area (middle) is inside the {C; N}-double-limitation region

*Figure 12: Single- and double-limitation growth regimes for N = 2 (carbon and nitrogen).*

In Figure 12, both axes are in log-scale, so any visually linear function is indeed linear. In chemostat cultures the volume is constant, so the uptake rate *q* of a substrate is proportional to the difference between its inlet and outlet concentrations. In the region to the left of the grey area, glucose (carbon, F2) is the limiting substrate as its outlet concentration close to zero, so $q_{F2}$ is always close to 100% of the supplied glucose.

If all substrates except F2 are in large excess and the biomass composition does not change significantly in this region, the growth rate is directly proportional to the F2 supply rate. This means the yield $Y_{X/F2}$ is constant. But if the biomass composition is constant, then consumption rate of ammonium (F1) is also directly proportional to the growth rate, so $Y_{X/F1}$ is also constant. This is equivalent to an elementary chemical reaction $\lambda_1$ F1 + $\lambda_2$ F2 → X, in which the $\lambda$ values are constant stoichiometric coefficients.

Thus, if the supply rates $c_{F1}$ and $c_{F2}$ are such that $c_{F2}/c_{F1} \leq \lambda_2/\lambda_1$, then $q_{F2} - c_{F2} \simeq 0$ and $c_{F1} - q_{F1} \simeq (\lambda_2/\lambda_1 - c_{F2}/c_{F1}) \cdot q_{F2}$. The left-hand side is the outlet concentration of ammonium and for a fixed $c_{F2} = q_{F2}$ (normalization), the right-hand side is a decreasing linear function of the glucose to ammonium ratio in the feed stream, exactly as depicted in Figure 12. This is what originates Definition 1 (single-limitation regimes).

If F2 is the most critical factor for the organism, then the average growth rate of its population will never be below the dashed curve in the plot of Figure 11b, simply because of selective pressure. Thus, the lower bound of the growth rate is not zero but a function of the set of factor availabilities. The set of specific uptake rates exhibited by a population is always inside the region of multiple-limitation (including its frontiers). Then, this region is used to demonstrate by absurd the validity of Principle II, as illustrated in Figure 13.

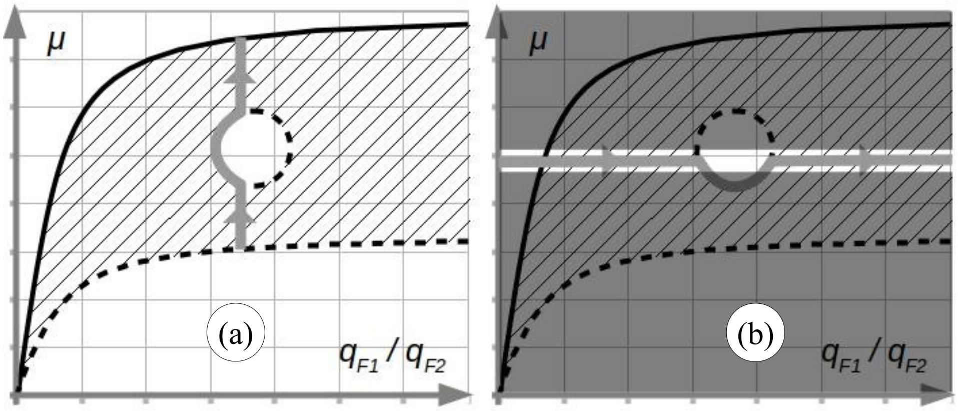


*Figure 13: Absurdity of a disjoint multiple-limitation region: a) the two objectives of maximizing the growth rate and minimizing substrate subutilization would be irreconcilable inside the Pareto-efficient region, what is a contradiction by definition; b) it would be impossible to reach a steady-state for some values of the specific growth rate set by the dilution rate of a continuous bioreactor.*

Therefore, Principle II implies the multiple-limitation region must be joint, but not convex, since there are many straight lines in Figure 13 that connect points near the origin to points near the dashed frontier without being entirely contained in the hatched region. The shape in Figure 11a is convex in 3D (the dashed curve is not behind but beneath the solid curve), however the non-linear transformation of it into the double-limitation region of Figure 13 does not preserve convexity (Lindahl, 2015).

The set of admissible flux distributions of a metabolic network is always a convex set (Wagner & Urbanczik, 2005) like the 3D Pareto region exemplified in Figure 11a. A multiple-limitation region like that of Figure 13 can be determined experimentally from single-limitation experiments (Zinn et al, 2004), which are unambiguous according to Definition 1. Thus, it is possible to validate the predictions of a metabolic network either by applying the non-linear transformation on them or the reverse transformation on the graph of the region.

### 5.1.2 The sigmoid effect of maintenance energy

So far, the three metabolic principles are being considered without the maintenance energy requirement, as it is visible in Figure 13, in which both saturation curves are Monod-like without an offset parameter like $a_H$ from the Herbert maintenance model (Equation 1). Figure 14 is Figure 11 with a positive $a_H$ for the dashed curve and a negative $a_H$ for the other. The negative $a_H$ is to model energy storages.

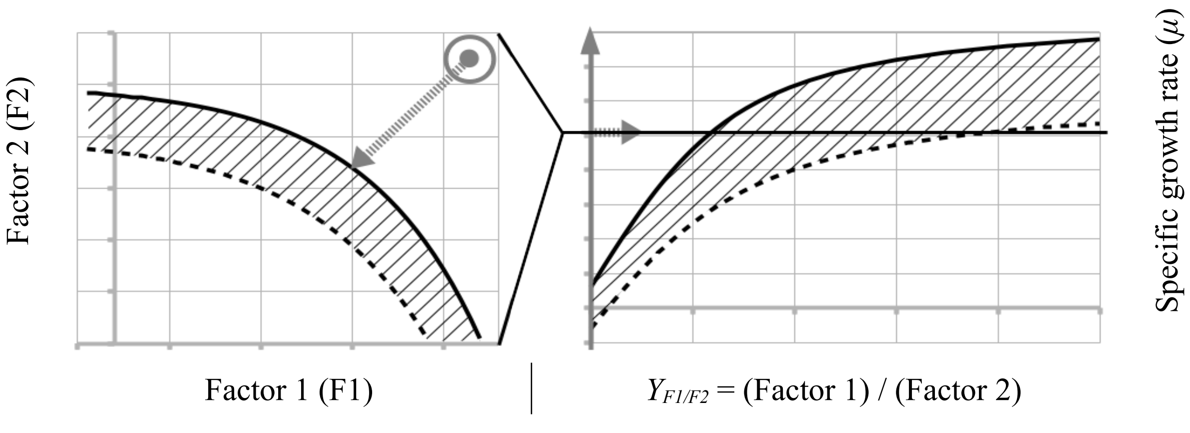
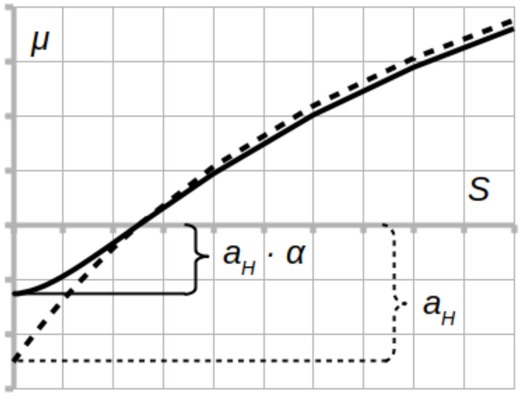


*Figure 14: Projection of two Pareto frontiers for the availabilities of two growth-limiting factors (a, adapted from Gorban et al, 2011) and its corresponding plot of the multi-objective specific growth rate in function of the ratio between the availabilities two factors (b). Added constant maintenance energy offsets ($a_H$).*

Without changing the values plotted as $\mu$, the effect of adding maintenance energy in the saturation curve is to displace the plot of the Pareto frontiers to the left generating negative values of availability of the normalized factor, as depicted in Figure 14a. One possible interpretation of a negative factor availability is that when a factor is unavailable in the environment, biomass may self-consume to obtain it. That is exactly the idea of "endogenous metabolism" behind the Herbert model, if F2 is the energy source (carbon).

According to section 3.3.2, it is known that the Herbert and Pirt models overestimate the maintenance energy requirement at $\mu < 0.1\ \mu_{max}$. Since low growth rates are difficult to achieve experimentally, it is possible that Monod-like kinetics on which these models are based is only accurate at $\mu > 0.1\ \mu_{max}$.

Replacing the Monod model by a sigmoid curve can solve the problem, because the sigmoid does not need to be so steep at low $\mu$ in order to fit the same data at high $\mu$. In 1987, Pirt proposed a correction factor $\alpha$ such that $1 - \alpha$ is the dormant fraction of the population (Pirt, 1987). This factor was estimated from experimental data and the graph of $\alpha(\mu)$ had the shape of a steep saturation curve. Applying this correction factor, the plot of the corrected Equation 1 indeed becomes a sigmoid (Figure 15).



Original (Pirt, 1965; Wang & Post, 2012):
$$\mu(S) \;=\; \mu_{max,H} \cdot S/(K_S + S) - a_H$$
Where S is the concentration of the limiting substrate (implicitly normalized for populations of the same size)

Corrected (Pirt, 1987; Wang & Post, 2012):
$$\mu(S,\alpha) \;=\; \alpha \cdot [\mu_{max,H} \cdot S/(K_S + S) - a_H]$$

*Figure 15: Comparison between the shapes of the original (dashed) and corrected (solid) Herbert/Pirt models.*

Reconciling this sigmoid shape with the Pareto frontiers for the availabilities of two growth-limiting factors, Figure 16 is obtained from Figure 14. Care must be taken when analysing Figure 16a because the order of the levels from top to bottom is not grey-dashed, black-dashed and black-solid but grey-dashed, black-solid and black-dashed, like the order in Figure 16b. It looks like but it is not a bivariate probability distribution density function because there may be two values for each bivariate point.

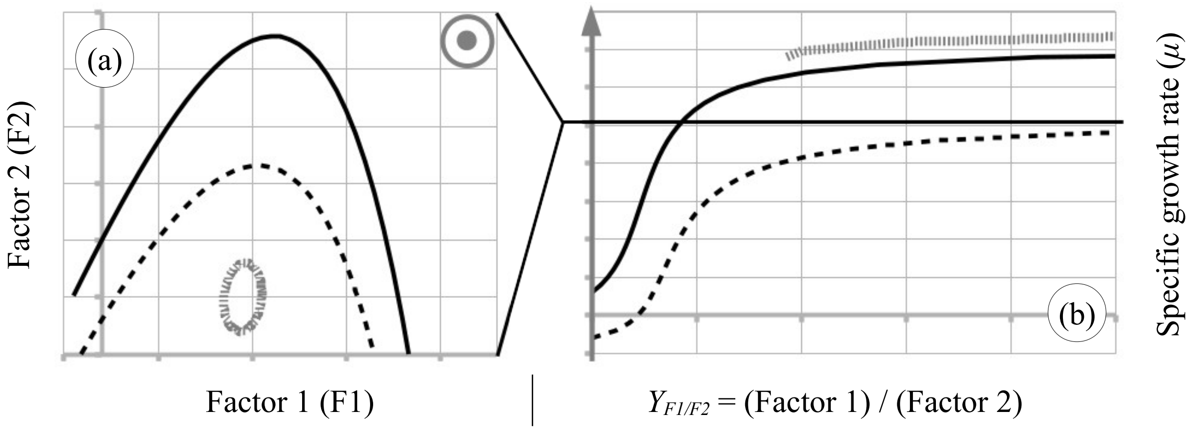
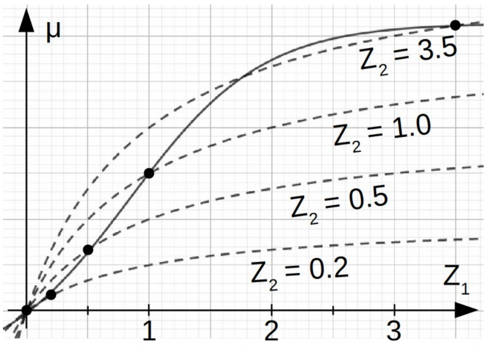


*Figure 16: Projection of two Pareto frontiers for the availabilities of two growth-limiting factors (a, adapted from Gorban et al, 2011) and its corresponding plot of the multi-objective specific growth rate in function of the ratio between the availabilities two factors (b). Added corrected variable maintenance energy model (α).*

In 3D, from top to bottom, the shape in Figure 16a first expands in space very fast, reaches a maximum perimeter and then shrinks slowly, like the tip of a Conical Asian Hat on top of a bucket. This convex 3D shape could be the intersection between a multiple of the joint probability distribution of the availabilities of F1 and F2 in the environment (the whole hat) and the convex cone of admissible flux distributions after adaptation (the bucket), an idea which is supported by recent theoretical developments (Gorban et al, 2016).

There are two more classes of evidences of this relationship between maintenance energy and a sigmoid saturation curve being a better model than Monod's. The first is depicted in Figure 17 (adapted from De Wit, 1994).



Adapted from Figure 4.3 of De Wit, 1994.
$Z_1$ and $Z_2$ are distinct growth factors, normalized for populations of the same size.

Each dashed curve is $\mu(Z_1)$ for a fixed value of $Z_2$. The solid curve is the projection of:

$$\mu = f(Z_1, Z_2) \cap \{Z_1 = Z_2\}$$

Onto the plane $\mu \times Z_1$

*Figure 17: The projection of $\mu(Z_1,Z_2)$ onto the plane $\mu \times Z_1$ is a sigmoid curve if $Z_1$ and $Z_2$ exert equal impact on $\mu$ and if both $\mu(Z_1)$ and $\mu(Z_2)$ are modelled by Monod-like curves.*

In Figure 17, maintenance energy is not considered, but adding the same offset $a_H$ in all Monod-like dashed curves would produce a sigmoid like the solid curve of Figure 15. This agrees with the direct normalization of Definition 3 if the reference substrate is the energy source, because both $\mu(Z_1)$ and $\mu(Z_2)$ would be offset by the same $a_H$.

The second class of extra evidence is the spatial position of the *banana*-shaped multiple-limitation regions in Egli space reported in the literature (Egli, 1991, Zinn et al, 2004; Taciro, 2008). The inverted normalization amplifies the horizontal distances (orthogonal to the $\mu$ axis) at low $\mu$ (low $Z_i$) and shrinks them at high $\mu$ (high $Z_i$). It also switches them (the points are mirrored around an axis parallel to $\mu$). However, the *bananas* are not very close to the $\mu$ axis, meaning that they cannot be constructed from Monod-like curves but can be from sigmoid curves. Figures 19 and 18 show examples respectively for $N = 3$ (3D) and $N = 2$ (2D).

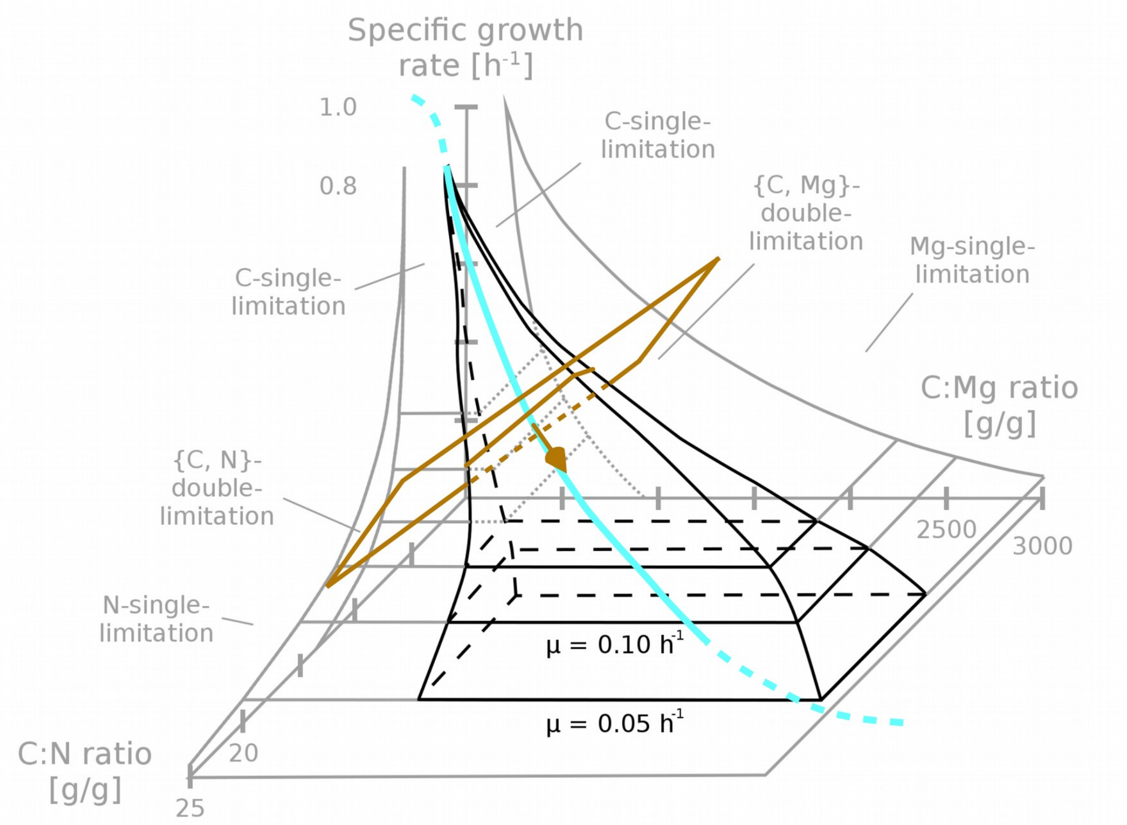
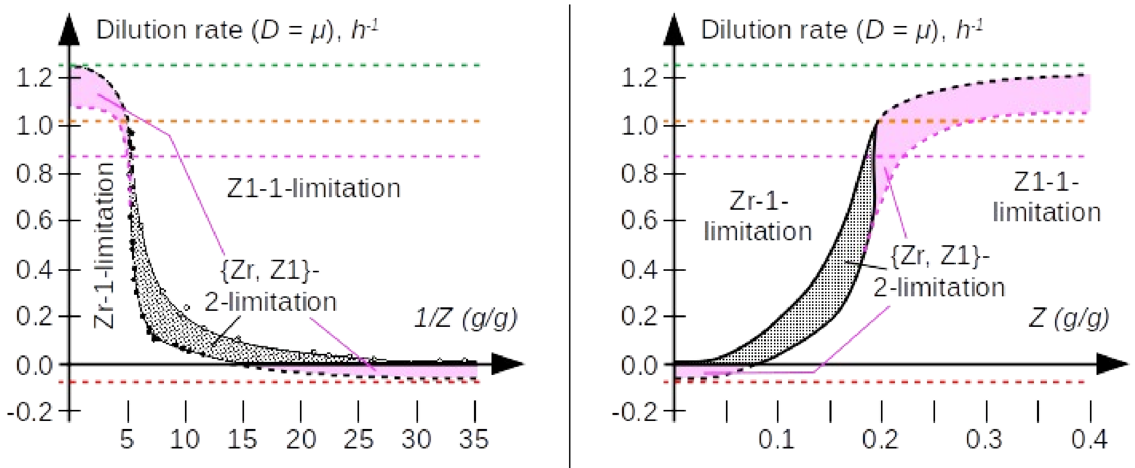


*Figure 18: Rectangular bounds of a 3D carbon, nitrogen, magnesium triple-limitation region in Egli space constructed from 2D double-limitation data. In greyscale, the adapted image from Egli, 1991. In colours, the light-blue sigmoid central behaviour and a brown cutting plane orthogonal to its tangent. The light-blue dashed segments are the effects of maintenance energy, which are more visible in Liebscher space.*



a) Egli space

b) Liebscher space

*Figure 19: Effect of inverting the normalization of the substrate concentration for a saturation curve of specific growth rate (μ). In greyscale, the data from Egli, 1991, originally in Egli space (a, inverted normalization). In colours, the features amplified by the direct normalization (b, direct normalization).*

Again, maintenance energy is not explicitly considered in these empirical studies, but their results agree with the inverted normalization of Definition 3 if the reference substrate is the energy source. It is the same rationale discussed in Figure 17.

This effect of maintenance energy is frequently neglected because most data in literature is from cultures under energy limitation, either caused by direct energy-source substrate limitation (e.g. carbon) like the black dashed curves of Figure 16 or by no substrate limitation like the grey dashed curves. In those conditions, maintenance is minimum or energy storages are available, masking the sigmoid behaviour.

Those two apparently opposite conditions are actually similar precisely because of the maintenance process. Maintenance consumes energy regardless of where it comes from (Zeroth Law of Thermodynamics), unlike biosynthesis which depends both on energy and on the chemical nature of substrates. And since maintenance is the most basic and prioritized process, without which nothing is possible because the organism dies, energy is always a much more critical factor than any chemical nature (type of mass).

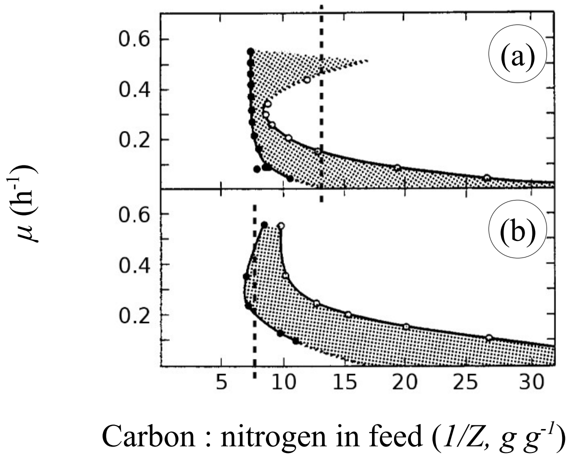
Thus, for chemoorganotrophs, the difference between $\mu_{max}$ in the *0-lim* regime and $\mu_{min}$ in a *{C, S}-[1+dim(S)]-lim* is whether carbon is limiting only due to energy or also due to its chemical nature. In other words, if biosynthesis competes with maintenance for energy. With the concept of quantitative limitation, these two apparent opposite conditions are respectively equivalent to “all factors are equally limiting including energy” and “all factors can be considered equally limiting relative to energy”. This is captured by choosing the energy source substrate as the reference substrate Sr.

The sequence of Figures 11, 14 and 16 form a possible evolutionary path of a metabolic network adjusting itself to the factor availabilities of the environment, which are also dynamic and affected by life itself (Figure 1). The sigmoid behaviour of the growth rate could be a consequence of the characteristic probability distribution of those availabilities.

The inflection point is the feature missing in the traditional maintenance energy models and in that point the substrate uptake metabolism overcomes the endogenous metabolism, which is then only used to survive under hostile conditions. Any saturation curve without an inflection point, like the Monod model, is a good approximation of the upper part of a sigmoid curve, which is the part most easily and frequently observed empirically.

### 5.1.3 Anomalous multiple-limitation regions

Anomalous double-limitation regions were reported (Egli, 1991). They are reproduced in Figure 20. The underlying phenomenon is also explained in that study: the higher the average carbon oxidation state in the carbon sources, the more the region is displaced to the right in Egli space, and also these organisms re-consume previously excreted substrates.



This Figure was adapted from Egli, 1991. The curves of the frontiers of the double-limitation region do not pass the vertical line test.

a: *Candida utilis* (glucose and $NH_4^+$)
b: *Candida valida* (ethanol and $NH_4^+$)

*Figure 20: Anomalous double-limitation regions.*

Once again, this is an effect of a complex energy management mechanism in the metabolism, like the endogenous metabolism. Both plots in Figure 20 suggest that at high growth rates the overflow metabolism takes over, secreting fermentation products that are later re-consumed. This is an apparent contradiction, because this metabolism is less efficient than the regular one, but it may be faster (Szenk et al, 2017).

This impossibility to avoid the concurrence of multiple substrates violates the hypothesis of substrate non-interchangeability from Definition 1. Still, these anomalous cases could be accounted for by correcting the uptake rates of carbon sources according to their average oxidation states.

A large part of the phenotypic data of *Pseudomonas* sp. LFM046 is from experiments with a mixture of glucose and fructose. Even though it is known they are metabolized by different pathways that likely generate different proportions of reducing power per mole of carbon (section 3.4.1), glucose and fructose have the same average carbon oxidation state.

### 5.1.4 Formalization

The objective of this section is to derive a mathematical description for the *N*-multiple-limitation region which represents the *N-lim* limitation regime of an organism. The regime *N-lim* can be determined from experimental data of *1-lim* regimes, which by definition can be unambiguously observed. Thus, the first formalization is what is a limitation regime and its particular case *S-1-lim* (Definition 1).

According to section 5.1.1, the three metabolic principles identified in literature together with the maintenance energy can be encoded as a sigmoid function for *μ(Z)*. Equation 25 is the generalization of the normalized uptake rate *Z* using Equation 7. Equation 26 is the generalization of the limiting substrate S to a set of concomitantly limiting substrates. Statements 27 to 29 are the definition of a *N*-D sigmoid manifold for $\mu$.

$$\vec{Z} = (Z_1, \ldots, Z_{N-1}) \ : \ Z_i = \begin{cases} q_{Si}/q_{Sr} & , \quad 1 \le i \le N \\ q_{Sr}/c_{Sr} & , \quad i = N \end{cases} \tag{25}$$

$$\begin{array}{lcc} S & = & \{S_i \ \forall i : S_i \text{ is limiting}\} \\ S^C & = & Sa \setminus S \end{array} \tag{26}$$

$$\mu = f_S(\vec{Z}) \ , \ f_S : \mathbb{R}_{>0}^{N-1} \to \mathbb{R} : f_S \in C^k(\mathbb{R}_{>0}^{N-1}) \ , \ k \ge 2 \tag{27}$$

$$\exists Z_{i,1}^* : \forall z \ge Z_{i,1}^* \ , \ 0 < \frac{df_S(z)}{dZ_i} \le \epsilon 1 \ , \ \forall i : S_i \in S^C \setminus S_r \ , \ \epsilon 1 \in \mathbb{R}_{>0} \tag{28}$$

$$\exists Z_{i,2}^* : \forall z \ge Z_{i,2}^* \ , \ -\epsilon 2 \le \frac{d^2 f_S(z)}{d{Z_i}^2} < 0 \ , \ \forall i : S_i \in S_a \ , \ \epsilon 2 \in \mathbb{R}_{>0} \tag{29}$$

Statement 27 simply is that $f_S$ is a sigmoid manifold taking *N-1* variables (the dimension of $\vec{Z}$) and is continuous at least up to its second derivative. Statement 28 defines a tolerance $\epsilon_1$ and states the first derivative of $\mathrm{f_S}$ is positive and lower than $\epsilon_1$ from some point onwards. And Statement 29 is the same but for the second derivative and $f_S$ being between a negative tolerance $-\epsilon_2$ and zero. This is the formalization of Principle I.

Due to Principle II, limitation is not a binary concept. A substrate can be continuously less, equally or more limiting than another. Then, there must be a tolerance $\epsilon_R$ to discriminate between *S* and $S^C$. Statement 28 cannot achieve this because if the mass fraction of $Z_i$ in biomass is higher than that of $Z_j$, it could be that $|d\mu / dZ_i| > \varepsilon_R > |d\mu / dZ_j|$ even if $Z_j$ is more limiting than $Z_i$. However, the variation of $\mu$ due to the amount of a consumed substrate can be compensated by normalizing $\mu$ with respect to that consumption (Statement 30). Then, the *S-lim* regime can be defined as a subset of Liebscher space (Statement 31; Liebscher space was defined in Definition 3).

$$\left|\frac{d(\mu/||\vec{Z}||)}{dZ_i}\right| > \epsilon_R > \left|\frac{d(\mu/||\vec{Z}||)}{dZ_j}\right| \Rightarrow \frac{d[f_S(z)/||\vec{Z}||]}{dZ_i} > \epsilon_R > \frac{d[f_S(z)/||\vec{Z}||]}{dZ_j} \tag{30}$$

$$P = (\vec{Z},\mu) \in S\text{-}lim \subset \mathbb{R}^N \Leftrightarrow \frac{d[f_S(\vec{Z})/||\vec{Z}||]}{dZ_i} < \epsilon_R\ ,\ i = 1\ldots(N-1) \tag{31}$$

Statement 31 is a generalization of the informal definition of single-limitation regime (Egli, 1991). There, $N = 2$ and thus $\vec{Z} = Z_1$ (a scalar). Like in Definition 1, it requires that all yields $Y_{X/Zi}$ are constant, that is, the magnitudes of their derivatives must be lower than a tolerance $\varepsilon_R$. And for $N = 2$ in particular, $Y_{X/Z1} = \mu/Z_1 = f_{S1}(Z_1)/Z_1$, which is exactly what the vectorial function inside the derivative operator in this Statement reduces to.

Having defined a point in a limitation regime, Principle II can be applied to formalize the concept presented in Figure 13. Given a point *P* belonging to a *Si-1-lim* like in Statement 31 and moving *P* along a line *Q*, there are only two alternatives (Statements 32 and 33).

$$S_i\text{-}1\text{-}lim \text{ unbounded in the direction of } Q \Leftrightarrow \text{st. 14 holds, } \forall P \tag{32}$$

$$\begin{gathered} \text{Otherwise} \Rightarrow \text{st. 14 replaced by 13} \Leftrightarrow \\ P \text{ enters } \Leftrightarrow S'\text{-}k\text{-}lim : S_i \in S' \wedge k \geq 2 \Leftarrow \\ \Leftarrow \nexists Q : \\ \{Q \cap S_i\text{-}k_i\text{-}lim \neq \varnothing\} \wedge \{Q \cap S_j\text{-}k_j\text{-}lim \neq \varnothing\} \wedge \{S_i\text{-}k_i\text{-}lim \cap S_j\text{-}k_j\text{-}lim = \varnothing\} \end{gathered} \tag{33}$$

Statement 32 is trivial: if a region is unbounded in a certain direction any point in that direction belongs to that region. The variables $Z_i$ are all positive and unbounded, and $\mu$ is bounded (upper-bounded by $\mu_{max}$ and lower-bounded by the maximum endogenous metabolism rate), so the Liebscher space is unbounded in all directions orthogonal to $\mu$. For $N = 2$, this is the horizontal direction in the positive sense.

The other alternative is when $Q$ is not in one of those unbounded directions. Then, Statement 33 is that no culture can transition from one regime to another in a linear path in Liebscher space if their sets of limiting substrates do not share at least one common substrate. For $N = 2$, this alternative is any direction with a vertical component.

In other words, Principle II guarantees that any point in Liebscher space belongs to a limitation regime and that the transition between any two limitation regimes is smooth. However, if any transition is smooth because there must always be at least one common substrate, *∅-0-lim* could not transition to any regime without contradicting Statement 33. This is resolved by assuming *∅-0-lim* is indistinguishable from *$S_a$-N-lim* when all substrates are equally limiting. In particular, when $\mu \rightarrow \mu_{max}, \varnothing$. Thus, *∅-0-lim* ⊂ *Sa-N-lim*.

In order to study how the transitions between different limitation regimes can be represented, let $H_p$ be a hyperplane with normal vector $H_{pn}$ and intersecting all the axes of Liebscher space in the first orthant (positive $Z_i$ and $\mu$). The brown plane in Figure 18 is an example of $H_p$, but in Egli space. Then, $H_{pn}$ cuts the limitation regimes and these intersections forms a collection $C$ of manifolds $C_i$ with the same properties of an $N$-Venn diagram (Bannier & Bodin, 2017). These properties are Statements 34 to 36.

$$\exists C \ \wedge \ dim(C) = N \tag{34}$$

$$\bigcap_{i=1}^{N} R_i \neq \varnothing \ , \ R_i = int(C_i) \vee R_i = ext(C_i) \tag{35}$$

$$dim(C_i \cap C_j) = n_{ij} \neq 0 : n_{ij} \text{ is finite} \ , \ \forall i \neq j \tag{36}$$

Statement 34 is trivial: there are $N$ single-limited regimes and each one of them corresponds to a $C_i$ manifold in the hyperplane $H_p$. Statements 36 and 35 respectively correspond to Statements 32 and 33. Since $Q$ is a line and $H_p$ is a *(N-1)*-D hyperplane, there exists always a $Q$ parallel to $H_{pn}$. If st. 32 is true for that $Q$, then st. 33 must be true for any $Q'$ orthogonal to $Q$, because $Q$ and $Q'$ are complementary as these st. 32 and 33. Then, $Q'$ is a line on the equivalent *N*-Venn diagram and thus satisfies st. 35 and 36. The last subcondition of st. 33 is the first subcondition of st. 36, as concluded in Statement 37.

$$\{\nexists Q' : dim(\{Q \cap C_i\} \cap \{Q \cap C_j\}) = 0\} \Leftrightarrow dim[C_i \cap C_j] \neq 0 \tag{37}$$

The finite $n_{ij}$ in Statement 36 merely excludes the possibility that some part of $C_i$ is tangent to some part of $C_j$, so $C_i$ and $C_j$ must necessarily cross each other. Since a $C_i$ corresponds to a *1-lim* regime, this means *Q'* can only cross two *S-1-lim* regimes without crossing also an *S'-k-lim* with $k \geq 2$ if $k = 0$. But $k = 0$ corresponds to the *0-lim* regime, which is a subset of the *N-lim* regime, so there are two consequences (Statements 38 and 39, $\forall i \neq j$).

$$\{\exists Q' : Q' \cap C_i \neq \varnothing \neq Q' \cap C_j \Leftarrow Q' \cap \varnothing\text{-}lim \neq \varnothing\} \Leftrightarrow \text{st. } 18 : R_i = ext(C_i) \tag{38}$$
$$\{\exists Q' : Q' \cap C_i \neq \varnothing \neq Q' \cap C_j \Leftarrow Q' \cap N\text{-}lim \neq \varnothing\} \Leftrightarrow \text{st. } 18 : R_i = int(C_i) \tag{39}$$

But st. 39 is the particular case of st. 33 for $k_i = k_j = 1$, because the *N-lim* regime contains all *N* substrates (and also the empty substrate set). Then, the condition *{Q' ∩ N-lim ≠ ∅}* in st. 39 is necessarily equivalent to condition *{$S_i$-1-lim ∩ $S_j$-1-lim ≠ ∅}* in st. 33, which in turn then correctly states *Q'* must exist. Since each curve $C_i$ of the *N*-Venn diagram may correspond to one and only one *S-1-lim* regime, this demonstrates that the manifolds formed in the hyperplane $H_p$ by the limitation regimes can be described by a *N*-Venn diagram.

### 5.1.5 Simplifying assumptions

In section 4.1.1, these four simplifying assumptions were briefly presented:

I. The ideal growth stoichiometry is constant (the sigmoid axis is coplanar with $\mu$);

II. The energy storage capacity is equal to the maximum maintenance energy;

III. The capacity to consume $S_i$ to save $S_j$ is equal to that of doing the opposite;

IV. The capacities in Assumptions II and III increase with the availability of the normalized factors ($||\vec{Z}||$) in the proportion of the ideal growth stoichiometry according to a Monod model (a saturation curve that starts almost linear).

Assumption I implies that the biomass elemental composition only varies if the proportion and/or elemental composition of the excreted products vary, due to mass conservation. This makes the central sigmoid *1*-D path be coplanar with the $\mu$ axis.

Assumption II and III are actually the same, but II is for the energy-source pseudo-substrate and III is for all other pseudo-substrates. A pseudo-substrate is equivalent to the set of all interchangeable substrates from which the population can harvest one chemical element. The energy-source substrate uptake rate is $q_C$ because *Pseudomonas* sp. LFM046 is a chemoorganotroph. The mixture of glucose and fructose gives $q_C = q_{C,glu} + q_{C,fru}$, where $q_{C,glu}$ and $q_{C,fru}$ are respectively the masses of carbon consumed from each of them.

In Assumption III, Si and Sj are chemical elements with $i \neq j$, and none of them is the energy-source Sr. Thus, Si and Sj are non-interchangeable. So an extra consumption of one to save the other is not due to replacing one by the other but due to storages and regulation of metabolic functions within a limited flexibility of the biomass composition (points around but not on the plane of the sigmoid path).

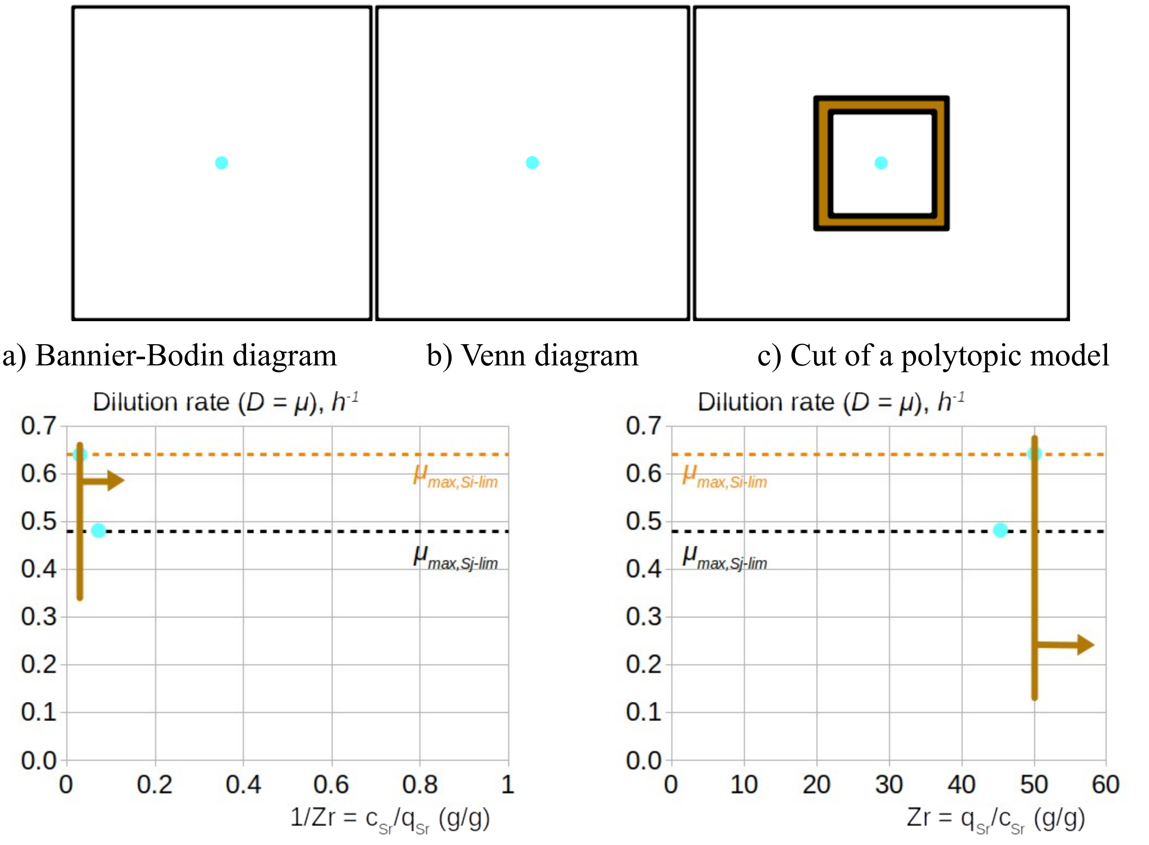
For example, P can be stored as polyphosphates when it is abundant and then consumed when it becomes scarce; when that happens, if N is abundant, $q_P/q_N$ will be lower than that corresponding to the growth stoichiometry and $q_N/q_P$ will be higher, so the global effect is that extra N is being consumed to save the scarce P. Assumption III simply means that in a comparable opposite situation of N scarcity and P abundance, the global effect is quantitatively the exact opposite. In other words, there is no hysteresis within a cycle of metabolic adaptation in less than one generation.

Analogously, in Assumption II, maintenance energy is a cost of opportunity for growth and measured as an equivalent negative growth rate. When this cost is higher than the growth rate itself, the endogenous metabolism is observed as a negative net growth rate. Thus, in terms of rates, Assumption II simply states that the maximum gain in $\mu$ achieved by consuming stored energy is equal to the maximum cost in μ due to the maintenance process.

Assumptions I to III make the *N*-phenotypic model much simpler and are not simply mathematical constructs, they are based on general reasonable assumptions of symmetry between the effect of biological processes of organisms selected over millions of years. Assumption IV is one way to implement saturation curves without an inflection point for the consumption of all chemical elements due to their chemical nature, including carbon.

### 5.1.6 Examples

Following the formalization from section 5.1.4, one way to describe the limitation regimes in Liebscher space is a *1*-D sigmoid path *PΣ* (St. 27 to 29) embedded in Liebscher space (St. 32) with a hyperplane $H_p$ orthogonal to this path moving along in this path. *PΣ* is entirely contained in the *N-lim* regime, which is surrounded by all other regimes filling the whole space. The cross-section imprinted on $H_p$ can be described by an *N*-Venn diagram. Figure 21 shows the case $N = 0$.



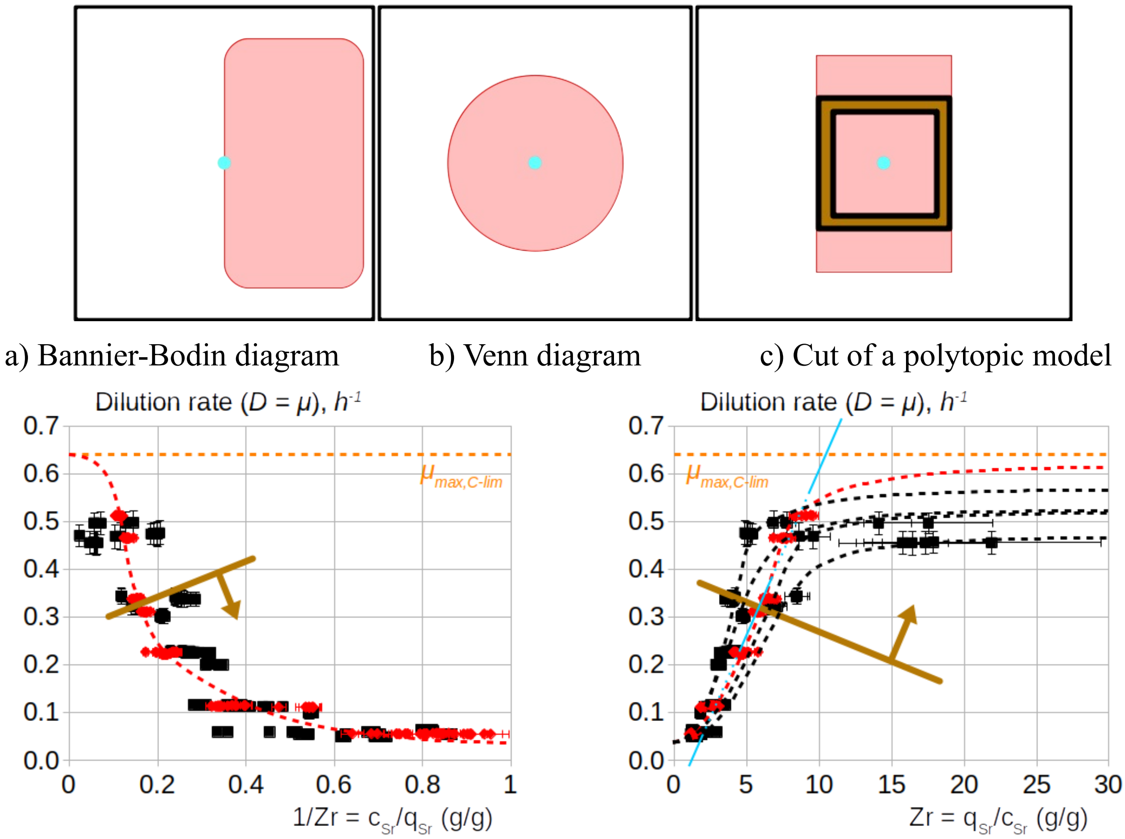
a) Bannier-Bodin diagram b) Venn diagram c) Cut of a polytopic model

d) Two distinct regimes in Egli space e) Two distinct regimes in Liebscher space

*Figure 21: Example of N-phenotypic model for N = 0 (d and e). Cross-section $H_p$ in brown (a, b and c).*

This case $N = 0$ is the *0*-data or binary data mentioned in section 3.2.1. It is simply a measurement of growth or non-growth, with no information on growth-limiting factors or the growth rate. It is a point lying either on the horizontal axis or anywhere else in space.

Figure 22 shows an example of $N = 1$. The red curve in item e is what the traditional maintenance energy models (Herbert and Pirt) approximate. If growth is limited by the energy substrate (Sr), $Z_r = q_{Sr}/c_{Sr} = 1$ regardless of $\mu$, but if $c_{Sr}$ is instead arbitrated from another fixed reference culture, it is obtained the curve $\mu = f_{Sr}(q_{Sr}/c_{Sr})$, which can be approximated by a Monod curve minus a constant $a_H$. That is the Herbert maintenance model (1964).



a) Bannier-Bodin diagram

b) Venn diagram

c) Cut of a polytopic model

d) Five distinct regimes in Egli space (data of *Pseudomonas* sp. LFM046, only the curve of *{Carbon}-k-lim*, $k \geq 1$, $c_{Sr} = 0.05$)

e) Item d in Liebscher space (data of *Pseudomonas* sp. LFM046)

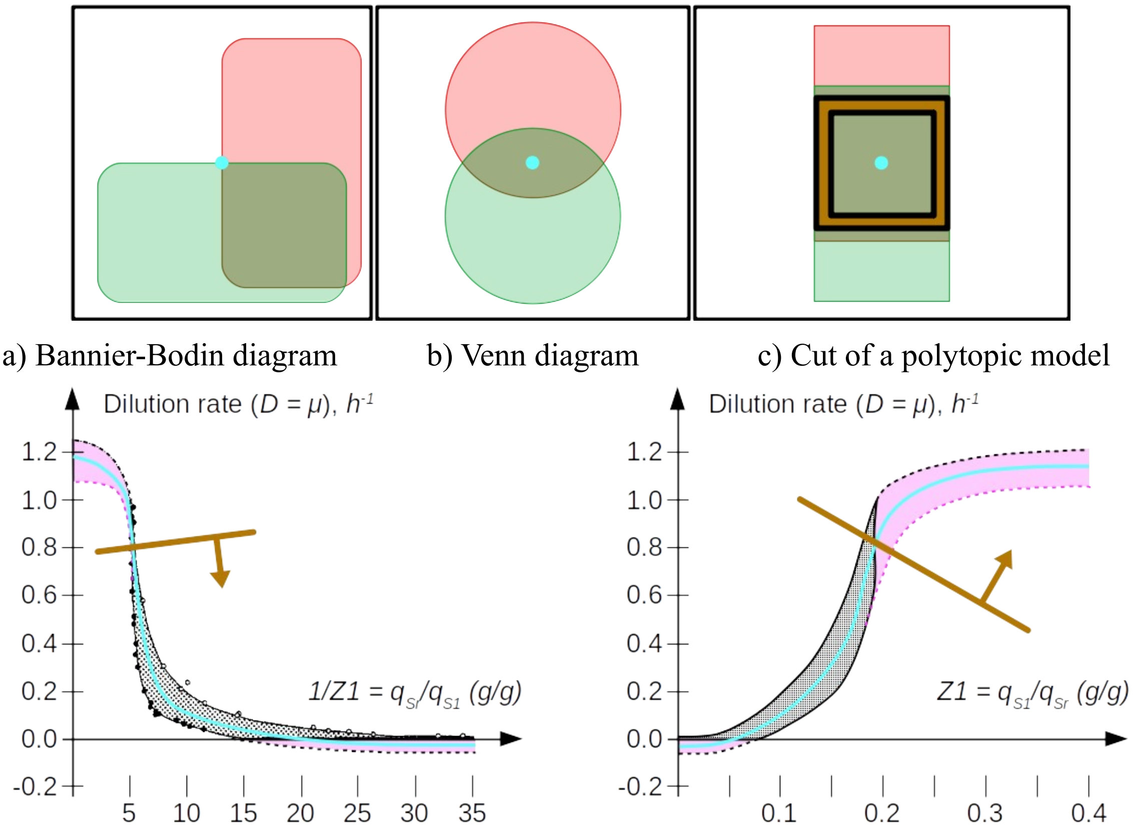
*Figure 22: Example of N-phenotypic model for $N = 1$ (d and e). Cross-section $H_p$ in brown (a, b and c).*

The Pirt model (1965) is the Herbert model with a change of variables (section 3.3.2): $a_H = Y_G\ m$ and $q_{Sr} = \mu/Y_G + m$, where $Y_G$ is the maximum theoretical yield, in this case, the maximum theoretical $Y_{X/Sr}$. In Figure 22e, it is visible that the light-blue line approaches the red curve for $0.05 < \mu < 0.5$ h$^{-1}$ (the interval $\mu > 0.1\ \mu_{max}$). This line is the Pirt model, only plotted with the axes switched (at $\mu = 0$, $Zr = m > 0$). And $c_{Sr}$ is contained inside $Y_{X/Sr}$.

Figure 23 shows an example for $N = 2$. Comparing to $N = 1$, it is clear that each curve in the previous case is actually two superimposed curves $f_K(Z) : K = S \vee K = S^C$, which have zero area in between them, for any $S$. In particular, for $S = S_a$, $f_K(Z)$ divides the space in two, naively like: $\{\mu < f_{Sa}(Z) \Leftrightarrow S_a\text{-}N\text{-}lim\} \wedge \{\mu \geq f_{Sa}(Z) \Leftrightarrow \varnothing\text{-}0\text{-}lim\}$. But since no culture can be outside all regimes and $\varnothing\text{-}0\text{-}lim \subset S_a\text{-}N\text{-}lim$, the partitioning must be like in Equation 40.

$$P = (\vec{Z}, \mu) \in K \ , \ K = \begin{cases} \varnothing\text{-}0\text{-}lim & , \quad \mu = f_{Sa}(\vec{Z}) \\ Sa\text{-}N\text{-}lim & , \quad \mu \neq f_{Sa}(\vec{Z}) \end{cases} \tag{40}$$

With these definitions, *0-lim* is not only when growth happens at the maximum growth rate, but when all substrates are equally limiting and thus are being supplied with the ideal proportion to the organism according to Principle III. This broader definition of “no limitation” is stoichiometric in nature, unlike the definition based on the growth rate.



a) Bannier-Bodin diagram

b) Venn diagram

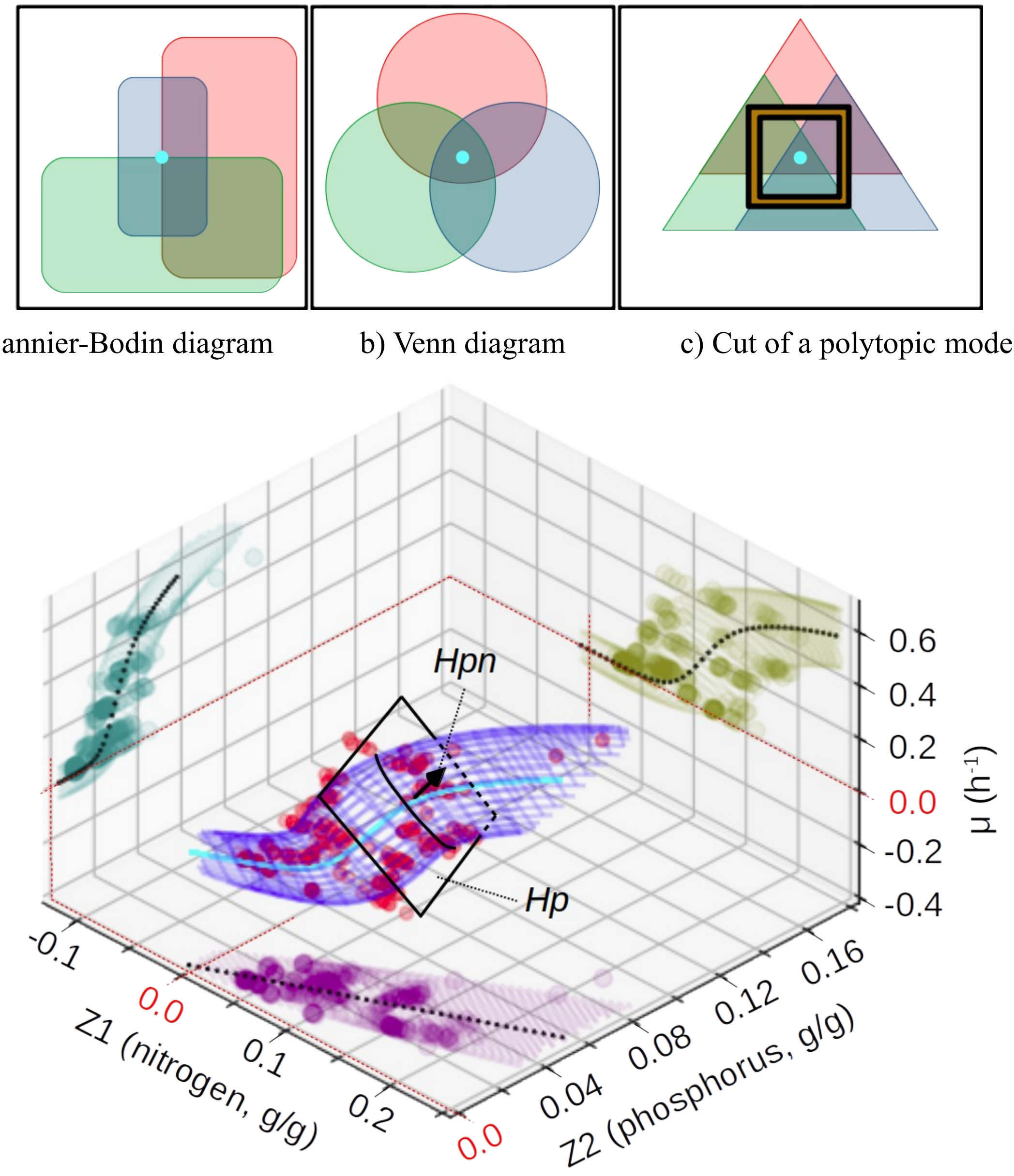
c) Cut of a polytopic model

d) *{C,N}-2-lim* in Egli space (data of *Klebisiella pneumoniae*, from Egli, 1991)

e) Item d in Liebscher space (OBS: S1 = nitrogen)

*Figure 23: Example of N-phenotypic model for N = 2 (d and e). Cross-section $H_p$ in brown (a, b and c).*

Figure 24 shows an example for *N* = *3*, only for the Liebscher space. An example of rectangular bounds in Egli space has been presented in Figure 18. Comparing all cases from 0 to 3, it is evident that case *N-1* is always a projection of case *N*. For example, the turquoise projection in Figure 24d is *{C,P}-2-lim*, with the same shape as Figure 23e. The purple projection is a transformation of the Pareto region of the nitrogen and phosphorus, like the two factors F1 and F2 in Figures 11, 14 and 16, but the two factors do not conflict (Figure 8).



a) Bannier-Bodin diagram

b) Venn diagram

c) Cut of a polytopic model

d) *{C,N,P}-3-lim* in Liebscher space (data of *Pseudomonas sp. LFM046*)

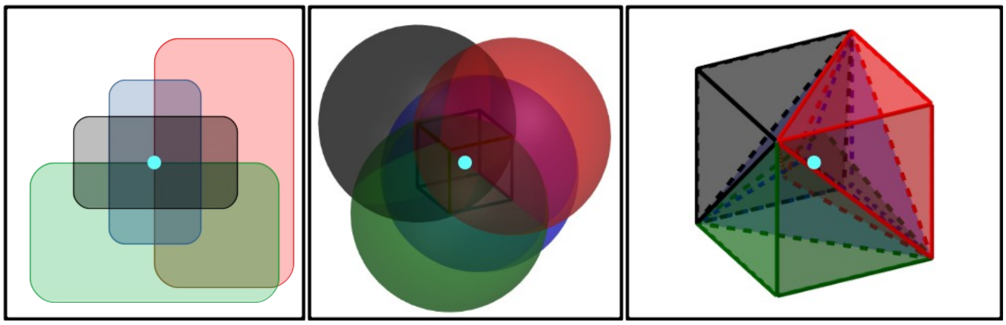
*Figure 24: Example of N-phenotypic model for N = 3 (d). Cross-section $H_p$ in black (a, b and c).*

From $N = 2$ to $N = 1$, the multiple-limitation region is projected in such a way that its upper and lower bounding curves align. So the only information in the $1$-D case is the core sigmoid behaviour of the specific growth rate in function of the normalized uptake rate of energy. This is why the case $N = 1$ can be used to estimate the maintenance energy requirement like the traditional maintenance models regardless of whether the experimental points are of energy-limited growth.

Each one of the five curves in Figure 22e ($N = 1$) is constructed from a dataset of a different limitation regime: *{N}-1-lim*, *{P}-1-lim*, *{N, P}-2-lim* and *{C, $S_u$}-k-lim*, where $S_u$ is an unknown set of substrates with $0 \leq dim(S_u) \leq dim(S_a)$. Only the two first regimes are of single-limitation, in which all points certainly belong to the same regime (Definition 1).

The last two are actually classes of regimes containing infinite degrees of limitation of each substrate sufficiently more limiting than the others (tolerance $\epsilon_R$ from Statement 31), so the experimental error in the points attributed to the same regime can bias the clustering of points into regimes. This is a consequence of the ambiguity of *S-k-lim* regimes with $k > 1$.

In light of the Equation 40, each regime k-lim with k ≤ N touches the N-lim from a different side (N-lim touches itself from all sides at once) and N-lim in turn surrounds 0-lim, so each curve in Figure 22e is a projection of a k-lim regime aligned with 0-lim. Inverting this rationale of projection to extrusion and assuming the shape of N-lim is such that its cross-sections are ($N$-$1$)-ellipsoids in Figure 24d (ellipses), it is obtained Figure 25 for $N = 4$.



a) Bannier-Bodin diagram b) Venn diagram c) Cut of a polytopic model

*Figure 25: Example of a cross-section of an N-phenotypic model for N = 4, represented in 3 ways.*

## 5.2 The genome-scale metabolic network

The latest version of the genome-scale metabolic network is v18, with 1485 reactions and 1257 metabolites (artificial reactions and metabolites included). The software *N*-GlobalFit was used between versions 12 and 13, only as a proof-of-concept compared to the ordinary manual method. The simulations to evaluate the primary and secondary hypotheses started in version 15. The log of the manual method up to version 12 can be found in Appendix A.

The final list of potential reactions used as input for the refining of the metabolic network had 1758 reactions (section 4.2.3) and the draft network generated from the genome in KBase had 1406 reactions. These two sets of reactions had 679 reactions in common, of which 246 were not exact matches (e.g. 1 $H^+$ of difference). The manual method up to v18 added 63 reactions to the draft, of which 34 were present in the list of potential reactions and the other 29 were manually searched in the METANETX and modelSEED databases.

Since this list was built from a gene orthology analysis comparing *Pseudomonas* sp. LFM46 to two well-studied similar *Pseudomonas* strains, these numbers imply that 48% of the draft network and 54% of the missing reactions could be predicted by gene orthology alone. This confirms that the manual method is a reliable unbiased control to assess the automated proof-of-concept *N*-GlobalFit, because the two percentages are similar.

However, it also shows that the method proposed in the publication of GlobalFit (Hartleb et al, 2016) was biased by the choice of organisms used to validate it as a method itself, as well as by the usual limitation regimes of the binary phenotypic data in the literature, because approximately half of the network would not be possible to determine for a truly novel organism using only the list of potential reactions, automatically or not.

The method used by the authors built lists of potential reactions for *E. coli* and *Mycoplasma genitalium* which were comprehensive enough because the reported GPR associations of these organisms are comprehensive and reliable enough, what is not the case of a novel organism not so similar with previously established ones. The same goes for the phenotypic data, which in the present case was not of carbon limitation or no limitation like are most of the cases reported in literature.

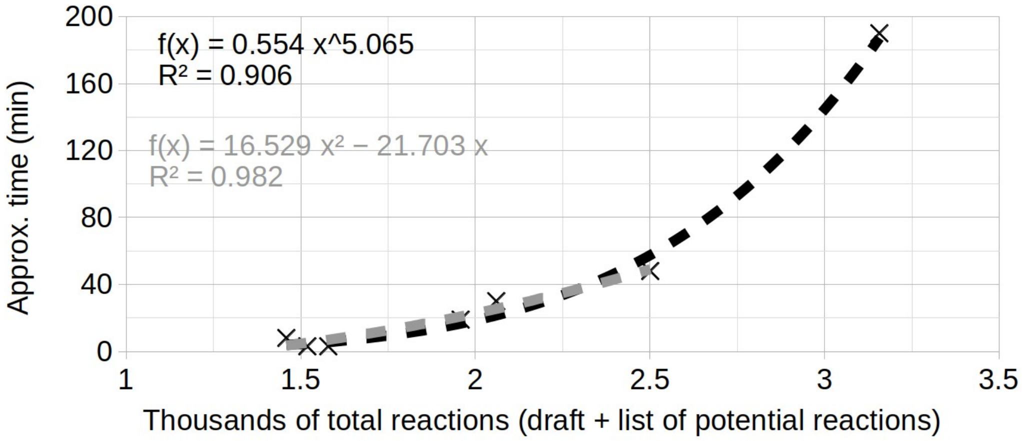
## 5.2.1 *N*-GlobalFit

The software *N*-GlobalFit was validated as a viable automated solution for refining genome-scale metabolic networks based on an *N*-phenotypic model adjusted to an *N*-database. Table 5.1 and Figure 26 summarize the results of the software tests.

*Table 5.1: Proof-of-concept of the automated metabolic network refining software N-GlobalFit*

| Test ID | Draft network | | Reaction database | | Points tested / predicted | Approx. time (min) |
|---|---|---|---|---|---|---|
| | ID | Reactions | ID | Reactions | | |
| draft | rem62 | 1397 | - | 0 | 00 / 16 | - |
| control | kbase12e3 | 1459 | - | 0 | 16 / 16 | 8 |
| rem07_t | rem07 | 1452 | trunc | 506 | 16 / 16 | 19 |
| rem02_m | rem02 | 1457 | manual3 | 122 | 16 / 16 | 3 |
| rem62_m | rem62 | 1397 | | | 16 / 16 | 3 |
| sk_a62 | skeleton | 133 | anti62_chk | 1927 | 00 / 16 | 30 |
| sk_mnx | skeleton | 133 | mnx | 88488 | 00 / 16 | > 120 |
| remm_s | remm | 1405 | super12b | 1752 | 0* / 16 | 190 |
| t2500a | rem62 | 1397 | manual3_t2500a | 1104 | 16 / 16 | 48 |
| t2500b | rem62 | 1397 | | | 00 / 21 | > 750 |

* 16 false positives (unrealistic high yields due to Energy-Generating Cycles)



| | |
|---|---|
| Hardware: | Intel Core i7-4610M @ 3.00 GHz ; 8 Gb DDR3 RAM @ 1.6 GHz |
| Software: | Debian 9.8 64-bit, Anaconda 3.10.5, r-base 3.5.1, r-cplexapi 1.3.3, r-globalfit 1.2, r-sybil 2.1.2, r-qclmatrix 0.97, r-irkernel 0.8.14 IBM CPLEX™ 12.8 |

*Figure 26: Computing time of N-GlobalFit in function of the total number of reactions to be tested for refining a genome-scale metabolic network of Pseudomonas sp. LFM046.*

According to Table 5.1 and Figure 26, there are two main factors that affect the application of the software *N*-GlobalFit: the total number of reactions (the draft network plus the list of potential reactions) and the existence of *N*-phenotypic points that are irreconcilable with this set of reactions (5 of the 21 points representative of the *3*-phenotypic model).

The complexity of the Mixed-Integer Linear Problem (MILP) solved by the software is in most cases dominated by the number of integer variables, which in this case-study correspond to the reactions added to or removed from the network. The irreconcilable points increase the number of possibilities that have to be tested before finding a solution, and if these points are made mandatory, no solution will be found at all after testing all possibilities.

Still, the results show that with a prior filtering step of removing inconsistent reactions and testing with subsets of the phenotypic points, the software can run in reasonable time on an ordinary machine. Considering that manually finding and resolving each inconsistency locally may take days of work from specialized researchers, this is a promising alternative.

The "skeleton" draft network was a biomass reaction plus exchange reactions (glucose, $NH_4^+$, $O_2$, $CO_2$, etc). The "mnx" list is the whole METANETX database with all reactions in both directions. It was not possible to build a metabolic network from scratch in reasonable time using only these inputs which require no extra information, but the results show that this may be possible (e.g. using a modified FastGapFilling algorithm).

As to the validation of the metabolic network of *Pseudomonas* sp. LFM046, Table 5.2 presents the 21 selected *3*-Liebscher points (the same from Figure 10), the predicted $\mu$ for each one of them (FBA, constraining $Z_1$ and $Z_2$) and the relative error ($\mu_{v12}/\mu - 1$).

*Table 5.2: Validation of the metabolic network (v12) with 16 points in N-GlobalFit (20% tol.)*

| ID | Shell surface | $Z_1$ (nitrogen) | $Z_2$ (phosphorus) | $\mu$ | $\mu_{v12}$ | Error |
|---|---|---|---|---|---|---|
| | | $(g\ g^{-1}\ h^{-1})\ (g\ g^{-1}\ h^{-1})^{-1}$ | | $g\ g^{-1}\ h^{-1}$ | | % |
| 20108 | upper | 0,05 | 0,01 | 0,06 | 0,08 | 19 |
| 30301 | upper | 0,13 | 0,04 | 0,47 | 0,41 | -13 |
| 40101 | upper | 0,08 | 0,01 | 0,12 | 0,08 | -29 |
| 40305 | upper | 0,1 | 0,02 | 0,3 | 0,23 | -23 |
| 50104 | upper | 0,07 | 0,02 | 0,23 | 0,15 | -35 |
| 50204 | upper | 0,08 | 0,03 | 0,34 | 0,23 | -32 |
| 50402 | upper | 0,12 | 0,05 | 0,47 | 0,3 | -35 |
| 120101 | upper | 0,11 | 0,01 | 0,06 | 0,05 | -17 |

| ID | Shell surface | $Z_1$ (nitrogen) | $Z_2$ (phosphorus) | $\mu$ | $\mu_{v12}$ | Error |
|---|---|---|---|---|---|---|
| | | (g g$^{-1}$ h$^{-1}$) (g g$^{-1}$ h$^{-1}$)$^{-1}$ | | g g$^{-1}$ h$^{-1}$ | | % |
| 130103 | upper | 0,03 | 0,01 | 0,07 | 0,06 | -15 |
| 140105 | upper | 0,19 | 0,03 | 0,51 | 0,54 | 4 |
| 150506 | upper | 0,19 | 0,02 | 0,35 | 0,28 | -20 |
| 20102 | lower | 0,06 | 0,01 | 0,06 | 0,08 | 25 |
| 40202 | lower | 0,09 | 0,02 | 0,23 | 0,17 | -24 |
| 50304 | lower | 0,04 | 0,02 | 0,05 | 0,04 | -20 |
| 50401 | lower | 0,16 | 0,07 | 0,49 | 0,32 | -36 |
| 50503 | lower | 0,05 | 0,03 | 0,12 | 0,07 | -40 |
| 80204 | lower | 0,19 | 0,04 | 0,32 | 0,36 | 12 |
| 140203 | lower | 0,16 | 0,02 | 0,1 | 0,08 | -18 |
| 140304 | lower | 0,15 | 0,02 | 0,06 | 0,06 | 0 |
| 160103 | lower | 0,13 | 0,04 | 0,23 | 0,19 | -15 |
| 170205 | lower | 0,13 | 0,03 | 0,06 | 0,05 | -5 |

In Table 5.2, the grey cells in the "Error" column are the set of 5 irreconcilable points. The error must be lower than +20% for the points on the upper surface of the 3D shell and higher than -20% for the points on the lower surface. The first point is within that tolerance but the solver could not find an optimal solution with it in reasonable time, meaning that it conflicts with the others even if it falls within the tolerance.

This point is near the upper surface of the shell and has the ID 20108, meaning that it is the observation #8 of the steady-state #201. But the point 20102, from the same steady-state, is near the lower surface. This is probably why one of them is irreconcilable with many points of the other surface (the set of irreconcilable points is naturally not unique).

These two points have very similar coordinates (same steady-state) but are near the origin, where all points inside the *N-lim* regime are close together. They are also the points of highest PHA content among the 21. Point #20108 is nearer the upper surface, where $\mu$ is higher and thus more difficult to achieve in a FBA simulation, so among the two, it was used for all simulation sets that required a fixed PHA content and/or a fixed specific growth rate.

## 5.2.2 The manual method

Table 5.3 presents the evolution of the predictive error with the version of the manually refined genome-scale metabolic network of *Pseudomonas* sp. LFM046. This error is the same as that of Table 5.2 ($\mu_{version}/\mu - 1$). The *N*-phenotypic model allowed an objective evaluation of the overall predictive quality of each network (grey rows in Table 5.3).

*Table 5.3: Predictive errors of the metabolic networks of Pseudomonas sp. LFM046*

| ID | Shell surface | Error (%) | | | | |
|---|---|---|---|---|---|---|
| | | v12 | v13 | v14 | v15 | v18 |
| 20108 | upper | 19 | 6 | 217 | 28 | - |
| 30301 | upper | -13 | 5 | 94 | 8 | -34 |
| 40101 | upper | -29 | -31 | 40 | -28 | -24 |
| 40305 | upper | -23 | -7 | 95 | -7 | -8 |
| 50104 | upper | -35 | -21 | 158 | -16 | -16 |
| 50204 | upper | -32 | -18 | 153 | -11 | -18 |
| 50402 | upper | -35 | -22 | 72 | -14 | -30 |
| 120101 | upper | -17 | -27 | 153 | -12 | -12 |
| 130103 | upper | -15 | -24 | 156 | -8 | -8 |
| 140105 | upper | 4 | -5 | 62 | 8 | -36 |
| 150506 | upper | -20 | -29 | 94 | -14 | -14 |
| 20102 | lower | 25 | 10 | 205 | 34 | 34 |
| 40202 | lower | -24 | -24 | 83 | -18 | -18 |
| 50304 | lower | -20 | -4 | 263 | 8 | 8 |
| 50401 | lower | -36 | -23 | 14 | -23 | -35 |
| 50503 | lower | -40 | -27 | 189 | -18 | -18 |
| 80204 | lower | 12 | 26 | 71 | 1 | -10 |
| 140203 | lower | -18 | -28 | 94 | -12 | -12 |
| 140304 | lower | 0 | -11 | 130 | 8 | 8 |
| 160103 | lower | -15 | 3 | 96 | 7 | 7 |
| 170205 | lower | -5 | 15 | 124 | 16 | 16 |
| Avg. abs. (wrong upper)* | | 11.5 | 5.50 | 118 | 14,7 | 0** |
| Avg. abs. (wrong lower)* | | 22.6 | 19,5 | 0 | 17,8 | 18,6 |

* Avg. abs.: average of the absolute values of the wrong predictions (> 0 for “upper”, < 0 for “lower”)
** Point #20108 is infeasible in version 18, so this average only considers the other 10 points

Version 12 had a fixed low PHA proportion of 0.30 moles per mole of biomass. That is equivalent to a PHA content of 4.7 %wt, just to verify the PHA synthesis was unblocked in the network. The proper PHA contents were included only in version 13, with a BIOTOT reaction for each one of the 21 points, with the PHA proportions varying from 0.01 to 8. And that is the only difference between versions 12 and 13.

PHA de-routes carbon and reducing power from growth. Therefore, intuitively, including PHA in the simulations should decrease the flux of biomass production, decreasing the predictive error on the upper surface of the 3D shell and increasing on the lower surface. However, both errors decreased. This indicates that in versions 12 and 13 there is an excess of reducing power that hinders growth if it is not discharged into something like PHA, as it has been reported for other *Pseudomonads* (Escapa et al, 2012; Nikel et al, 2015).

In version 14, the alternative biomass reactions to test the secondary hypothesis (section 4.3.2) were unblocked. This greatly increased all biomass production fluxes, indicating that in the previous versions this flux was limited by the uptakes of nitrogen and phosphorus for all points and not just the ones in regimes *{N,P}-k-lim* with $k \leq 2$.

In version 15, the $CO_2$ production rate was constrained to its actual observed value for each point and that made all biomass production fluxes decrease again. Intuitively, this would be expected if the $CO_2$ production was underestimated by version 14, what would indicate the network had an unrealistically high overall energetic efficiency which could be corrected by a constraint of maintenance energy like a non-zero flux through an ATP sink.

However, in version 14 there was no $CO_2$ production at all. Thus, there is at least one Energy-Generating Cycle (EGC) in the network. An EGC is formed when two or more reactions have stoichiometries and fluxes such that their global reaction has only external metabolites (those that need not be balanced) and at least one conserved moiety in the direction of energy production, for example: $2\ H_2CO_3 + ADP \rightarrow 2\ H_2O + 2\ CO_2 + ATP$.

Indeed, over 100 reactions in versions 14 and 15 have fluxes near or at their artificial bounds, which are absolute values orders of magnitude larger than the ones set as actual constraints, only to make the optimization problem bounded for the solver. For example, all forced fluxes are in the range -10 to 10, the artificial bounds are -1000 to 1000 and some reactions have fluxes like -1000 and 995. An EGC is always going to be activated when maximizing the biomass, so reactions with fluxes close to the artificial bounds are candidates for participants of an EGC.

An EGC is often caused by reactions in wrong directions, like reversible reactions that should actually be irreversible. There may be more than 50 EGC's in version 14, and that can be used favourably: PHA accumulation is induced under energy excess and while EGC's are relatively easy to detect and solve, structural problems that decrease the overall availability of energy in the network like a metabolite with two different ID's are not.

Thus, some EGC's were kept if they were not associated with another kind of problem and the simulations to evaluate the primary and secondary hypotheses were formulated considering that the network has EGC's. This is particularly important for the secondary hypothesis, as discussed in section 5.3. From version 15 to 18, three EGC's were found:

I. Phosphoenolpyruvate/oxaloacetate;

II. NAD/NADP transhydrogenases, with the reactions:

- rxn00083: $NAD^+ + NADPH \leftrightarrow NADH + NADP^+$ ,
- rxn10125: $NADH + NADP^+ + 2\ H^+_e \leftrightarrow NAD^+ + NADPH + 2\ H^+$ .

III. Malonyl-CoA production, with the reactions:

- rxn06672: $ATP + H_2CO_3 + \text{cofactor1} \leftrightarrow ADP + PO_4^{3-} + H^+ + \text{cofactor2}$ ,
- rxn06673: Acetyl-CoA + cofactor2 ↔ Malonyl-CoA + cofactor1 ,
- other reactions.

The workaround for EGC I was to make the reactions irreversible in the direction of consuming phosphoenolpyruvate, as reactions like these were in the small-scale model (Taciro, 2008). EGC II was trivial to find and solve, yet very important for the proposed simulations, for which none of the NAD/NADP transhydrogenase reactions can be reversible.

EGC III was found during these simulations because it was being used not only to produce ATP but also Malonyl-CoA, which is a necessary precursor for the fatty-acids *de novo* synthesis. This was an example of an EGC associated with another structural problem, which in this case was the lack of a production route for Malonyl-CoA (R_rxn00258_c0).

## 5.3 Hypotheses and simulations

For the primary hypothesis, it was defined the flux $v_{ff}$, equivalent to the total turnover of the fatty-acids futile cycle: $v_{ff} = v_{R\_rxn05343\_c0} - v_{R\_rxn00178\_c0}$. Reaction R_rxn05343_c0 is one of the reactions between $C_8$ and $C_{10}$ in the fatty-acids *de novo* synthesis. Reaction R_rxn00178_c0 is the final reaction in β-oxidation, which produces Acetyl-CoA. If there is such futile cycle, $v_{ff}$ must be greater than zero even with:

- zero uptake rate of substrates metabolized by the β-oxidation, e.g. octadecanoate ($C_{18}$);
- a PHA composition poor in $HA_{MCL}$, e.g. 95 %mol of $C_4$, which then is necessarily de-routed from the *de novo* synthesis since β-oxidation is not the metabolization pathway;
- Constrained uptake fluxes of glucose, fructose, $NH_4^+$, $PO_4^{3-}$ and $O_2$ and production flux of $CO_2$, in proportions similar to those of point #20108.

Therefore, FBA simulations were carried-out by constraining the flux of the artificial reaction A_BIOTOT01_u to different values and maximizing $v_{R\_rxn05343\_c0}$ to calculate $v_{ff}$. This simulates the maximum turnover of the hypothetical fatty-acids futile cycle in function of the growth rate, always with the same high PHA content (approx. 55 %wt) and a non-naturally occurring composition (95 %mol of 3HB and the rest of $HA_{MCL}$). The result is Figure 27.

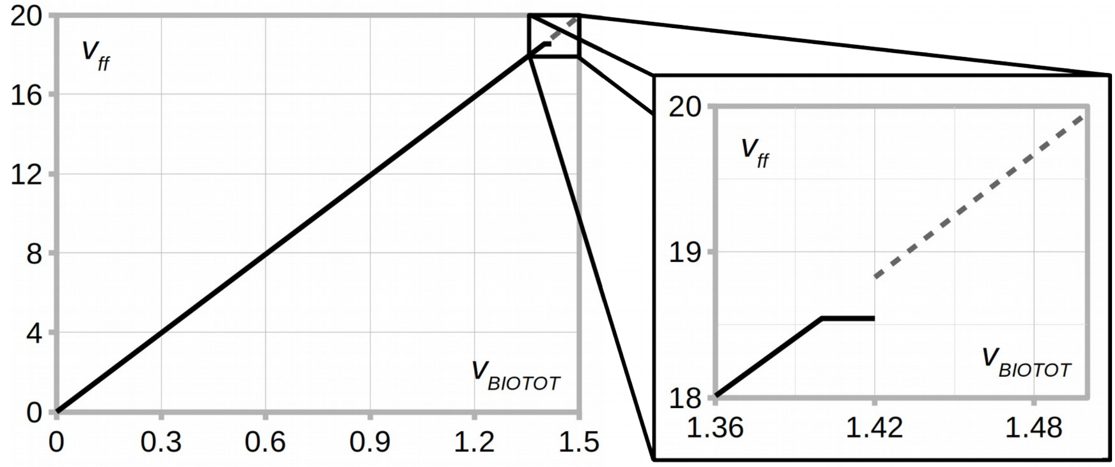


*Figure 27: Maximum turnover of the fatty-acids futile cycle ($v_{ff}$) in function of the growth rate with ca. 55 %wt CDW of a PHA with 95 %mol of 3HB ($v_{BIOTOT}$) in a metabolic network of Pseudomonas sp. LFM046. At vBIOTOT = 1.42, the simulation becomes infeasible (black solid line) because there phosphorus uptake is too low (this constraint is relaxed in the grey dashed line).*

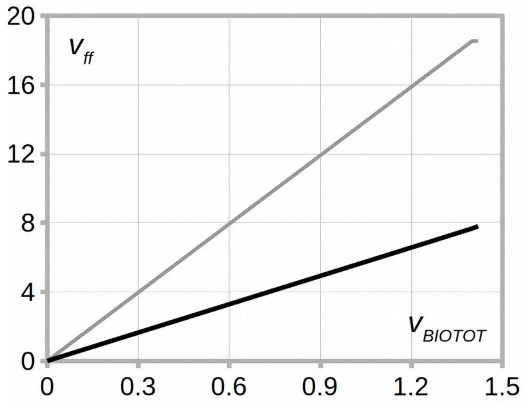
In Figure 27, $v_{ff}$ is always positive regardless of $v_{BIOTOT}$, what demonstrates the metabolic network of *Pseudomonas* sp. LFM046 is compatible with the hypothesized futile cycle. In other words, it is not suggested that the cycle *does* happen in the recombinant strain harbouring the PHA polymerase from *Aeromonas* sp., but that it *can* happen. Moreover, $v_{ff}$ increases with $v_{BIOTOT}$, meaning that the more PHA with this composition is produced, more intense the cycle can be.

The zoomed-in chart to the right shows that a plateau of $v_{ff}$ is reached, and after that a small increase in $v_{BIOTOT}$ makes the simulation infeasible. The grey dashed line that continues without this are simulations with a higher upper bound of phosphorus uptake. This indicates two important results:

I. the network predicts that $v_{ff}$ and by extension the whole metabolism are limited in phosphorus but not in carbon nor nitrogen, like indeed was the culture #20108;

II. this relationship between $v_{ff}$ and $v_{BIOTOT}$ is an upper bound estimate that follows Liebig's Law of the Minimum: $v_{ff}$ is either limited only by $v_{BIOTOT}$ or only by something else, without smoothly transitioning across multiple limitation regimes.

Figure 28 extends Figure 27 with the native PHA composition of *Pseudomonas* sp. LFM046, which has a much lower average carbon oxidation state.



Grey: PHA with 95 %mol of $C_4$ (3HB)

Black: PHA with the composition of the wildtype strain *Pseudomonas* sp. LFM046:
0.00 %mol of $C_4$
2.90 %mol of $C_6$
23.4 %mol of $C_8$
60.9 %mol of $C_{10}$
9.10 %mol of $C_{12\Delta5}$
3.70 %mol of $C_{12}$

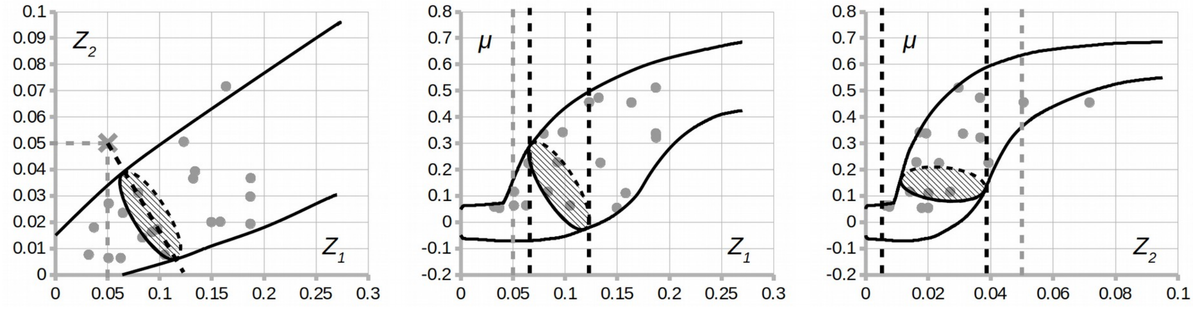
OBS: only the grey line reaches a plateau before simulations become infeasible

*Figure 28: Maximum turnover of the fatty-acids futile cycle ($v_{ff}$) in function of the growth rate with ca. 55 %wt CDW of PHA ($v_{BIOTOT}$) with two distinct compositions in a metabolic network of Pseudomonas sp. LFM046.*

Figure 28 shows that in the wildtype strain *Pseudomonas* sp. LFM046, $v_{ff}$ is less than half of that of the simulated strain for the same value of $v_{BIOTOT}$. And, as such, no plateau is reached before $v_{BIOTOT}$ is too high to keep the simulations feasible, implying that:

III. the wildtype strain *Pseudomonas* sp. LFM046 cannot dissipate as much energy in the fatty-acids futile cycle as the hypothetical simulated strain similar to the recombinant harbouring the PHA polymerase from *Aeromonas* sp, because the native PHA polymerase consumes $C_{10}$ faster removing more of it from the futile cycle;

IV. keeping at least one possible Energy-Generating Cycle (EGC) in the network is a way to separate the evaluation of the primary and secondary hypotheses from the evaluation of the boundaries of the limitation regimes with carbon excess, in which PHA accumulation is actually observed. This is because a fixed direction of $\vec{Z}$ crosses distinct limitation regimes depending on the value of $\mu$ (Figure 29).



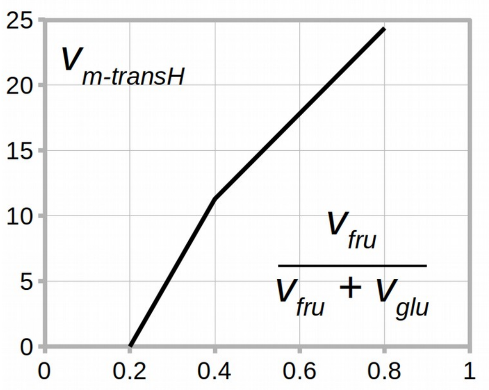
a) $Z_1 \times Z_2$ b) $Z_1 \times \mu$ c) $Z_2 \times \mu$

$Z_1$ and $Z_2$: nitrogen and phosphorus normalized uptake rates [(g g$^{-1}$ h$^{-1}$) (g g$^{-1}$ h$^{-1}$)$^{-1}$]
$\mu$: specific growth rate (h$^{-1}$)

*Figure 29: A point representing the conditions of a culture with phosphorus excess (grey cross) and a region describing possible specific growth rates for that condition (hatched area). The points are the 21 representative points of the boundaries of the {C, N, P}-3-lim regime of Pseudomonas sp. LFM046 (solid curves).*

Result III confirms the primary hypothesis and together with result I it suggests that the metabolic network is also compatible with the secondary. For the secondary hypothesis, the resulting plot is Figure 30, which shows that at least for most of the possible range of fructose proportion in the total carbon uptake, a solution using the transhydrogenase to accommodate the variation in the fluxes of NAD$^+$, NADH, NADP$^+$ and NADPH caused by the fatty-acids futile cycle is optimal.

There may be many equally optimal solutions. Constraining $v_{m\text{-}transH}$ to zero did not change the values of the objective function $v_{fru}$ + $v_{glu}$, confirming that the membrane-bound transhydrogenase is not the only way to accommodate fatty-acids futile cycle. Also, Figure 30 is not similar to Liebig's Law of the Minimum like the local upper bound obtained in Figure 27, but similar to a sigmoid like a global behaviour of Liebscher's Law of the Optimum.



PHA with 95 %mol of $C_4$

Constant fluxes:
Biomass production, PHA production

Constrained fluxes (uptake or production):
$NH_4^+$, $PO_4^{3-}$, $O_2$, $CO_2$

Objective function: minimize the sum of glucose and fructose uptake rates $v_{fru} + v_{glu}$

*Figure 30: Flux of the membrane-bound transhydrogenase in function of the fructose relative uptake.*

In Figure 30, $v_{m\text{-}transH}$ increases with the proportion of fructose. The futile cycle consumes NADPH and produces NADH, the opposite of the membrane-bound transhydrogenase, and fructose metabolization produces less NADPH per mole of carbon than that of glucose. Thus, the more fructose in relation to glucose, the less total availability of NADPH in the metabolism. Then:

V. the fatty-acids futile cycle or equivalents are part of the flux distributions that minimize the global consumption of the energy source, contrary to intuition;

VI. the metabolic network of *Pseudomonas* sp. LFM046 is compatible with the effect of replacing glucose by fructose as the energy source in the metabolism being similar to that of replacing phosphorus by nitrogen as the PHA accumulation inducer. Such effect would be the decrease of the excess of NADPH. This because:

- The constraints used for Figure 30 are those of phosphorus limitation (A_BIOTOT01_u and the uptake rates), which according to result I cause the network to be limited in phosphorus but not on nitrogen nor carbon,

- $v_{m\text{-}transH}$ increases in function of the fructose proportion in the carbon uptake, what is likely due to an increase of $v_{ff}$ or equivalents according to result V,

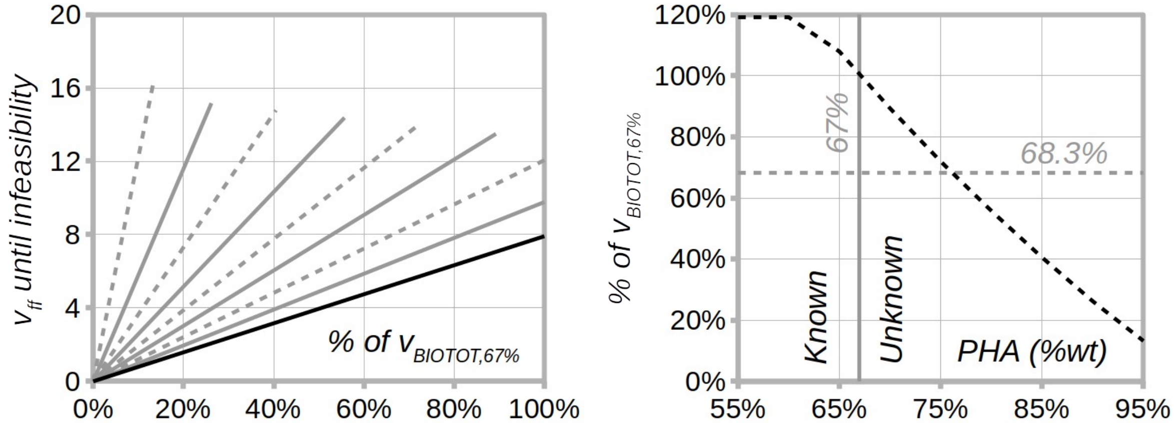
- $v_{ff}$ increases in function of $v_{BIOTOT}$ because of excess of reducing power equivalents (NADPH) and much too slow PHA polymerization, according to result III.

Result VI shows once again why keeping EGC's was important, since not doing so would not exclude carbon (energy) from being one of the possibly limiting substrates.

## 5.4 PHA bioprocess optimization strategies

According to Table 3.1, the PHA content increases with the average carbon oxidation state of the PHA. For pure PHA$_{SCL}$, there are multiple reports of content higher than 90% wt, and since the viable maximum is certainly lower than 100% wt, a realistic estimate is 95% wt. For PHA$_{MCL}$, the highest reported contents are around 67% wt (the database used here; Poblete-Castro et al, 2014). A realistic estimate was obtained by assuming that (Figure 31):

- the genome-scale metabolic network of the strain LFM046 represents well a metabolism which achieves this maximum PHA$_{MCL}$ content;
- the probability of a value of PHA$_{MCL}$ content being possible is 1 if this value is less than or equal to 67 %wt (because that was observed) and it decreases linearly with the maximum feasible flux through the fatty-acids futile cycle ($v_{ff}$);
- a realistic estimate corresponds to the probability of 68.3% (1 standard-deviation);



a) $v_{ff}$ in function of the normalized $v_{BIOTOT}$ black line is point #20108 (~55% PHA), grey lines are increments of 5% of PHA content (right to left)

b) highest values of $v_{ff}$ of item a in function of their corresponding PHA contents

*Figure 31: A realistic estimate of the maximum viable PHA$_{MCL}$ content based on the genome-scale metabolic model of Pseudomonas sp. LFM046 is 75 %wt (b, where the dashed lines meet)*

Thus, the maximum viable PHA content is estimated to range from 75 to 95 %wt. Results suggest it is a function of the average carbon oxidation state in the PHA (that is, the monomeric composition) because there is a permanent excess of reducing power in the central metabolism, the storage and dissipation capacities of it are a bottleneck during PHA accumulation and selective pressure favours the PHA composition which best fits the most frequent excess pool of reducing power, in a trade-off with other metabolic bottlenecks.

The idea that the maximum PHA content is not limited by the supply of PHA precursors but by other systemic factors like the storage capacity of reducing power or even mechanical stress agrees with recent improvements on $PHA_{SCL}$ content on *E. coli* not due to precursor supply but due to the shape of the cell (Chen & Jiang, 2017).

The relationship between shape and tolerance to mechanical stress is direct. The relationship between shape and the capacity to store or dissipate reducing power is indirect, and some examples are the hypotheses of actively fluidizing the cytoplasm to avoid molecular crowding (Fernandez-de-Cossio-Diaz & Vazquez, 2018) and of molecular crowding in the cell membrane where the electron transport chain is (Szenk et al, 2017).

For *Pseudomonads*, a whole body of evidence indicates their central metabolisms have a permanent excess of reducing power, possibly to neutralize oxidative stress. This favours the production of $PHA_{MCL}$ and requires specific strategies of reduction or dissipation of this excess for less reduced copolymer compositions, because the PHA polymerases with high affinity for $HA_{SCL}$ monomers are likely too slow compared to the fatty-acids metabolism.

This agrees with the observation that the wildtype strain LFM046 produces PHA with constant composition whereas the recombinant strain LFM461 *phaPCJ* seems to change its PHA composition depending on the time of cultivation in fed-batch and/or on the limiting nutrient(s) that are inducing PHA accumulation. In the first case, the PHA polymerase was selected over many generations together with the central metabolism, so it is well fit to the particular size of the pool of the reducing power equivalents. In the second, it was not.

Thus, the efficient control of the composition without substrate co-feeding requires manipulation of the relative speeds of the fatty-acids *de novo* synthesis and a recombinant PHA polymerase with monomer affinities complementary to the native one. The higher the speed of the *de novo* synthesis relative to that of the polymerase, the higher the proportion of $HA_{MCL}$ monomers. If this relative speed can be controlled on-the-fly during the bioprocess, it would be possible to control even the tertiary and quaternary structures of the PHA composition.

The results in section 5.3 showed that this relative speed can be controlled to some extent by choosing an appropriate limiting nutrient as PHA accumulation inducer and the carbon source: nitrogen and fructose slow down the *de novo* synthesis whereas phosphorus and glucose speed it up. To assess that together with the data already available, the next experiment should be with the recombinant strain using fructose as the sole carbon source and in the *N-1-lim* regime.

## 5.5 A note on maintenance energy

According to section 3.3.2, the most usual way of modelling maintenance is to constrain the flux of an artificial ATP sink reaction such that it brings down the maximum biomass yield from the energy source substrate (carbon) to an observed value. Then this flux is considered constant in order to predict flux distributions for conditions not yet observed. This is an application of the original Pirt or Herbert models of constant maintenance energy. Naturally, since this is based on the biomass yield from the energy source predicted by the network, it requires that all Energy-Generating Cycles (EGC's) are resolved before.

With the *N*-phenotypic model, this real biomass yield can be estimated independently of the metabolic network itself, using only previous measurements from a few cultures (e.g. 5 extreme points with $N = 3$). This allows the estimation of a value of maintenance energy for each predicted phenotype. It also allows two alternative methods to model maintenance energy which can account for EGC's, without the need to previously eliminate them:

- biomass sink: instead of consuming ATP, maintenance consumes biomass like a cost of opportunity for growth. The flux of this biomass sink is the difference between the maximum $\mu$ obtained by FBA and the $\mu$ predicted by the *N*-phenotypic model, for each point in *N*-Liebscher space. This forces the metabolic network to activate all pathways linked to biomass production even if the net $\mu = 0$. It is directly applicable and generates more realistic results for conditions like PHA fed-batch cultures in the accumulation phase;

- Energy-Dissipating Cycles (EDC's): the opposite of EGC's, like for example the very fatty-acids futile cycle proposed and analysed in this work. This is a concept even closer to what reality may be than the biomass sink. It can be quantitatively estimated the same way. The disadvantage is that it is not centralized in one single flux.

The quantitative estimation of maintenance energy with the *N*-phenotypic model does not depend on assumptions like it being constant or a linear function of $\mu$. A specific value can be estimated for each possible set of uptake rates (points in *N*-lim, refer to Figure 29).

## 5.6 A note on biomass composition

Genome-scale metabolic models usually have one single equation representing the biomass composition, based on the macromolecular composition of *E. coli* measured in cultures of carbon limitation or no limitation (Neidhardt et al, 1990). This simplification may indeed be valid in those limitation regimes (Puchalka et al, 2008; Van Duuren et al, 2013; Appendix A), much like the assumption of constant maintenance energy (section 5.5).

However, it is known that the biomass composition changes with the limitation regime and impacts the central metabolism (Taymaz-Nikerel et al, 2010; Dikicioglu et al, 2015; Folsom & Carlson, 2015). That has been confirmed for *Pseudomonas* sp. LFM046 by the results of section 5.2.2 (network v14) and 5.3 (secondary hypothesis). PHA accumulation is induced by nitrogen and phosphorus limitation in different and non-additive ways: the regime *{N,P}-2-lim* does not behave like a simple superposition of *N-1-lim* and *P-1-lim*.

In fact, these results indicate that considering a single biomass equation was actually one of the main problems in the previous attempt to build a large-scale metabolic model for *Pseudomonas* sp. LFM046 from the small-scale model (Taciro, 2008), in order to explain why PHA accumulation was not the same for all limitation regimes with carbon excess (section 3.4.1). Therefore, it is now clear that:

- metabolic models for bioprocess R&D cannot rely on a single biomass equation, specially if their corresponding organisms are to be used as bioplatforms whose potential for applications must be evaluated before defining any specific application;
- alternative biomass equations can be generated by matching their stoichiometries to alternative elemental compositions (C, H, N, O, P and S), which are much simpler to measure than macromolecular compositions (aminoacids, RNA, etc);
- building a genome-scale metabolic network by removing parts of a supernetwork is a better approach than adding parts to a subnetwork, because in the first case no sequence of removals needs to be defined (Acuña et al, 2009) and in the second the lack of an alternative biomass composition can bias the whole building process.

# 6 CONCLUSION

This work was divided in three steps: bioreactor data treatment and selection; construction and validation of a genome-scale predictive metabolic model for *Pseudomonas* sp. LFM046 using the bioreactor data; and *in silico* experiments using that model in order to propose strategies for the maximization of the intracellular polyhydroxyalkanoate (PHA) content with the control of its monomeric composition.

In the first step, it was developed the *N*-phenotypic model, which predicts all possible growth rates of an organism from measurements of *N* uptake rates of non-interchangeable substrates of as little as 5 cultures in distinct conditions. It integrates maintenance energy avoiding the main shortcomings of the traditional Pirt and Herbert models: overestimation at low specific growth rate and wrong predictions for growth under energy excess.

The *N*-phenotypic model was applied for the chemostat culture database of *Pseudomonas* sp. LFM046 with $N = 3$. It allowed the detection of outliers (about 10% of the initial 258 points) and the selection of 21 points which best represent the likely universe of all phenotypes of this strain.

For the second step, two methods were carried-out in parallel: the traditional manual method and the automatic *N*-GlobalFit software. The former was systematized and logged to provide an extensive example for future reference. The latter was validated as a proof-of-concept considering the resources usually available for refining a genome-scale metabolic network (time, computers and knowledge). Both methods were shown to be interchangeable and to benefit from the outlier filtering and point selection of the previous step.

In the third step, based on several independent experimental results, two hypotheses on the metabolism of PHA of *Pseudomonas* sp. LFM046 were formulated: a futile cycle of fatty-acids biosynthesis and degradation is used to dissipate a permanent excess of reducing power and under phosphorus limitation this mechanism is intensified. They were confirmed *in silico* by specific FBA simulations, showing that the PHA content can be maximized with the control of the monomeric composition by combining genetic modifications (PHA polymerase) with convenient bioprocess conditions.

# REFERENCES

Acuña V, Chierichetti F, Lacroix V, Marchetti-Spaccamela A, Sagot M-F, Stougie L. (2009). Modes and cuts in metabolic networks: complexity and algorithms. Biosystems; 95(1): 51-60. https://doi.org/10.1016/j.biosystems.2008.06.015

Acuña V, Milreu PV, Cottret L, Marchetti-Spaccamela A, Stougie L, Sagot M-F, (2012a). Algorithms and complexity of enumerating minimal precursor sets in genome-wide metabolic networks. Bioinformatics; 28(19): 2474-83. https://doi.org/10.1093/bioinformatics/bts423

Acuña V, Birmelé E, Cottret L, Crescenzi P, Jourdan F, Lacroix V, et al. (2012b). Telling stories: enumerating maximal directed acyclic graphs with a constrained set of sources and targets. Theoretical Computer Science; 457: 1-9. https://doi.org/10.1016/j.tcs.2012.07.023

Adami C. (2012). The use of information theory in evolutionary biology. Annals of the New York Academy of Sciences; 1256: 49-65. https://doi.org/10.1111/j.1749-6632.2011.06422.x

Altaee N, El-Hiti GA, Fahdil A, Sudesh K, Yousif E. (2016). Biodegradation of different formulations of polyhydroxybutyrate films in soil. SpringerPlus; 5(1): 762. https://doi.org/10.1186/s40064-016-2480-2

Anderson AJ, Dawes EA. (1990). Occurrence, metabolism, metabolic role, and industrial uses of bacterial polyhidroxyalkanoates. Microbiological Reviews; 54(4): 450-472. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC372789/pdf/microrev00039-0130.pdf

Andrade R, Wannagat M, Klein CC, Acuña V, Marchetti-Spaccamela A, Milreu PV, et al. (2016). Enumeration of minimal stoichiometric precursor sets in metabolic networks. Algorithms for Molecular Biology; 11: 25. https://doi.org/10.1186/s13015-016-0087-3

Bannier A, Bodin N. (2017). A new drawing for simple Venn diagrams based on algebraic construction. Journal of Computational Geometry; 8(1): 153-173. http://dx.doi.org/10.20382/jocg.v8i1a8

Baroukh C, Muñoz-Tamayo R, Steyer JP, 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. https://doi.org/10.1371/journal.pone.0104499

Bartell JA, Blazier AS, Yen P, Thøgersen JC, Jelsbak L, Goldberg JB, et al. (2017). Reconstruction of the metabolic network of *Pseudomonas aeruginosa* to interrogate virulence factor synthesis. Nature Communications; 8: 14631. https://doi.org/10.1038/ncomms14631

Besser J, Carleton HA, Gerner-Smidt P, Lindsey RL, Trees E. (2018). Next-generation sequencing technologies and their application to the study and control of bacterial infections. Clinical microbiology and infection; 24(4): 335-341. https://doi.org/10.1016/j.cmi.2017.10.013

BIOLOG. (2018). New metabolic models validated with phenotype microarrays [internet]. Hayward: BIOLOG. 2018 Dec. 13. Available from: https://biolog.com/2018/12/13/new-metabolic-models-validated-with-phenotype-microarrays/

Blumenstein K, Macaya-Sanz D, Martín JA, Albrectsen BR, Witzell J. (2015). Phenotype MicroArrays as a complementary tool to next generation sequencing for characterization of tree endophytes. Frontiers in microbiology; 6: 1033. https://doi.org/10.3389/fmicb.2015.01033

Bobadilla-Fazzini RA, Pérez A, Gautier V, Jordan H, Parada P. (2017). Primary copper sulfides bioleaching vs. chloride leaching: advantages and drawbacks. Hydrometallurgy 168; 26-31. https://doi.org/10.1016/j.hydromet.2016.08.008

Bolger M. (2018). Will bioplastics repeat the biofuels saga?. [opinion article]. Euractiv. 2018 Feb. 27. https://www.euractiv.com/section/agriculture-food/opinion/will-bioplastics-repeat-the-biofuels-saga/

Bordbar A, Yurkovich JT, Paglia G, Rolfsson O, Sigurjónsson ÓE, Palsson BØ. (2017). Elucidating dynamic metabolic physiology through network integration of quantitative time-course metabolomics. Scientific Reports; 7: 46249. https://doi.org/10.1038/srep46249

Bosi E, Donati B, Galardini M, Brunetti S, Sagot M.-F, Liò P, et al. (2015). MeDuSa: a multi-draft based scaffolder. Bioinformatics; 31(15): 2443–51. https://doi.org/10.1093/bioinformatics/btv171

Braunstein A, Muntoni AP, Pagnani A. (2017). An analytic approximation of the feasible space of metabolic networks. Nature Communications; 8: 14915. https://doi.org/10.1038/ncomms14915

Bresler G, Bresler M, Tse D. (2013). Optimal assembly for high throughput shotgun sequencing. BMC bioinformatics; 14(Suppl. 5): S18. https://doi.org/10.1186/1471-2105-14-S5-S18

Broddrick, J. (2018). Available predictive genome-scale metabolic network reconstructions. Available from: http://systemsbiology.ucsd.edu:80/InSilicoOrganisms/OtherOrganisms

Budinich M, Bourdon J, Larhlimi A, Eveillard D. (2017). A multi-objective constraint-based approach for modeling genome-scale microbial ecosystems. PLOS ONE; 12(2): e0171744. https://doi.org/10.1371/journal.pone.0171744

Bugge MM, Hansen T, Klitkou A. (2016). What is the bioeconomy? A Review of the Literature. Sustainability 2016; 8(7): 691. https://doi.org/10.3390/su8070691

Bugnicourt E, Cinelli P, Lazzeri A, Alvarez V. (2014). Polyhydroxyalkanoate (PHA): review of synthesis, characteristics, processing and potential applications in packaging. eXPRESS Polymer Letters; 8(11): 791–808. https://doi.org/10.3144/expresspolymlett.2014.82

Burgard AP, Nikolaev EV, Schilling CH, Maranas CD. (2004). Flux coupling analysis of genome-scale metabolic network reconstructions. Genome research; 14(2): 301–312. https://doi.org/10.1101/gr.1926504

Carcano S. (2010). A model for cell growth in batch bioreactors. [dissertation]. Milano: Politecnico de Milano; 2010. Available from: https://www.politesi.polimi.it/bitstream/10589/2082/1/2010_07_Carcano.pdf

Cardinali-Rezende J, Alexandrino PMR, Nahat RATPSN, Sant'Ana DPV, Silva LF, Gomez JGC, et al. (2015). Draft genome sequence of *Pseudomonas* sp. strain LFM046, a producer of medium-chain-length polyhydroxyalkanoate. Genome Announcements; 3: 45–46. https://doi.org/10.1128/genomeA.00966-15

Castro REN, Alves RMB, Nascimento CAO, Giudici R. (2018). Assessment of sugarcane-based ethanol production, fuel ethanol production from sugarcane. In: Basso TB, Basso LC, editors. IntechOpen. https://doi.org/10.5772/intechopen.78301

Cespedes LG, Nahat RATPS, Mendonça TT, Tavares RR, Oliveira-Filho ER, Silva LF, et al. (2018). A non-naturally-occurring P(3HB-*co*-3HA$_{\mathrm{MCL}}$) is produced by recombinant *Pseudomonas* sp. from an unrelated carbon source. International Journal of Biological Macromolecules; 114: 512–519. https://doi.org/10.1016/j.ijbiomac.2018.03.051

Chatterjee A, Huma B, Shaw R, Kundu S. (2017). Reconstruction of *Oryza sativa indica* genome scale metabolic model and its responses to varying RuBisCO activity, light intensity, and enzymatic cost conditions. Frontiers in plant science; 8: 2060. https://doi.org/10.3389/fpls.2017.02060

Chen G-Q, Jiang XR. (2017). Engineering bacteria for enhanced polyhydroxyalkanoates (PHA) biosynthesis. Synthetic and systems biotechnology; 2(3): 192–197. https://doi.org/10.1016/j.synbio.2017.09.001

Claudel-Renard C, Chevalet C, Faraut T, Kahn D. (2003). Enzyme-specific profiles for genome annotation: PRIAM. Nucleic Acids Research; 31(22): 6633–6639. https://doi.org/10.1093/nar/gkg847

Cosentino S, Iwasaki W. (2019). SonicParanoid: fast, accurate and easy orthology inference, Bioinformatics; 35(1): 149–151. https://doi.org/10.1093/bioinformatics/bty631

De Wit CT. (1994). Resource use analysis in agriculture: a struggle for interdisciplinarity. In: Fresco LO, Stroosnijder L, Bouma J, Van Keulen H, editors. The future of the land: mobilizing and integrating knowledge for land use options. Chichester: John Wiley & Sons, 1994. p. 41–55. http://library.wur.nl/WebQuery/wurpubs/24825

Devoid S, Overbeek R, DeJongh M, Vonstein V, Best AA, Henry C. (2013). Automated genome annotation and metabolic model reconstruction in the SEED and Model SEED. In: Alper H, editor. Systems Metabolic Engineering. Methods in Molecular Biology (Methods and Protocols). Totowa: Humana Press. 985 vol. https://doi.org/10.1007/978-1-62703-299-5_2

Di Genova A, Ruz GA, Sagot M-F, Maass A. (2018). Fast-SG: an alignment-free algorithm for hybrid assembly. GigaScience; 7(5): giy048. https://doi.org/10.1093/gigascience/giy048

Dikicioglu D, Kırdar B, Oliver SG. (2015). Biomass composition: the "elephant in the room" of metabolic modelling. Metabolomics; 11(6): 1690–1701. https://doi.org/10.1007/s11306-015-0819-2

Diniz SC, Taciro MK, Gomez JGC, da Cruz Pradella JG. (2004). High-cell-density cultivation of *Pseudomonas putida* IPT 046 and medium-chain-length polyhydroxyalkanoate production from sugarcane carbohydrates. Applied Biochemistry and Biotechnology; 119(1): 51–69. https://doi.org/10.1385/ABAB:119:1:51

Egli T. (1991). On multiple-nutrient-limited growth of microorganisms, with special reference to dual limitation by carbon and nitrogen substrates. Antonie van Leeuwenhoek; 60(3–4): 225–234. https://doi.org/10.1007/bf00430367

Ernsting A, Smolke R. (2018). Biofuelwatch report: dead end road - the false promise of cellulosic biofuels. London, Biofuelwatch; 2018. https://www.biofuelwatch.org.uk/wp-content/uploads/Cellulosic-biofuels-report-2.pdf

Escapa IF, García JL, Bühler B, Blank LM, Pietro MA. (2012). The polyhydroxyalkanoate metabolism controls carbon and energy spillage in *Pseudomonas putida*. Environmental Microbiology; 14(4): 1049–1063. https://doi.org/10.1111/j.1462-2920.2011.02684.x

Faria JP, Rocha M, Rocha I, Henry CS. (2018). Methods for automated genome-scale metabolic model reconstruction. Biochemical Society Transactions; 46(4): 931–936. https://doi.org/10.1042/BST20170246

Fell DA, Small JR. (1986). Fat synthesis in adipose tissue – an examination of stoichiometric constraints. The Biochemical Journal; 238(3): 781–786. https://doi.org/10.1042/bj2380781

Fernandez-de-Cossio-Diaz J, Vazquez A. (2018). A physical model of cell metabolism. Scientific Reports; 8(1): 8349. https://doi.org/10.1038/s41598-018-26724-7

Ferrarini MG, Siqueira FM, Mucha SG, Palama TL, Jobard É, Elena-Herrmann B, et al. (2016). Insights on the virulence of swine respiratory tract mycoplasmas through genome-scale metabolic modeling. BMC genomics; 17: 353. https://doi.org/10.1186/s12864-016-2644-z

Fritzemeier CJ, Hartleb D, Szappanos B, Papp B, Lercher MJ. (2017). Erroneous energy-generating cycles in published genome scale metabolic networks: identification and removal. PLOS Computational Biology; 13(4): e1005494. https://doi.org/10.1371/journal.pcbi.1005494

Folsom JP, Carlson RP. (2015). Physiological, biomass elemental composition and proteomic analyses of *Escherichia coli* ammonium-limited chemostat growth, and comparison with iron-

and glucose-limited chemostat growth. Microbiology; 161(8): 1659–1670. https://doi.org/10.1099/mic.0.000118

Gilbert D, Heiner M, Jayaweera Y, Rohr C. (2017). Towards dynamic genome-scale models. Briefings in Bioinformatics: bbx096 https://doi.org/10.1093/bib/bbx096

Golden JS, Handfield R, Pascual-Gonzalez J, Agsten B, Brennan T, Khan L, et al. (2018). Indicators of the U.S. biobased economy. Washington D.C., U.S. Department of Agriculture; 2018. https://www.usda.gov/oce/energy/files/BIOINDICATORS.pdf

Gomes RS. (2009). Obtenção de mutantes deficientes no acúmulo de PHA e construção de linhagens recombinantes para o controle da composição monomérica. [Ph. D. thesis (Biotechnology)]. Sâo Paulo, Universidade de São Paulo; 2009. Available from: http://www.teses.usp.br/teses/disponiveis/87/87131/tde-29042010-102817/pt-br.php

Gomez JGC. (1994). Isolamento e caracterização de bacterias produtoras de polihidroxialcanoatos. [Masters thesis (Biotechnology)]. São Paulo, Universidade de São Paulo; 1994.

Gomez JGC. (2000). Produção por *Pseudomonas* sp. de polihidroxialcanoatos contendo monômeros de cadeia média a partir de carboidratos: avaliação da eficiência, modificação da composição e obtenção de mutantes. [Ph. D. thesis (Microbiology)]. São Paulo: Instituto de Ciências Biomédicas, Universidade de São Paulo; 2000.

Gorban AN, Pokidysheva LI, Smirnova EV, Tyukina TA. (2011). Law of the minimum paradoxes. Bulletin of Mathematical Biology; 73(9): 2013–2044. https://doi.org/10.1007/s11538-010-9597-1

Gorban AN, Tyukina TA, Smirnova EV, Pokidysheva, LI (2016). Evolution of adaptation mechanisms: adaptation energy, stress, and oscillating death. Journal of Theoretical Biology; 405: 127–139. https://doi.org/10.1016/j.jtbi.2015.12.017

Greene, J. (2012). PLA and PHA biodegradation in the marine environment. [Contractor's report]. Chico, Chico Research Foundation & California State University; 2012.

Groot WJ, Sikkenk CM, Waldram RH, Van der Lans RGJM, Luyben KChAM. (1992). Kinetics of ethanol production by baker's yeast in an integrated process of fermentation and microfiltration. Bioprocess Engineering; 8(1–2): 39–47. https://doi.org/10.1007/BF00369262

Harrison JP, Boardman C, O'Callaghan K, Delort A-M, Song J. (2018). Biodegradability standards for carrier bags and plastic films in aquatic environments: a critical review. Royal Society Open Science; 5(5). http://doi.org/10.1098/rsos.171792

Hartleb D, Jarre F, Lercher MJ. (2016). Improved metabolic models for *E. coli* and *Mycoplasma genitalium* from globalfit, an algorithm that simultaneously matches growth and non-growth data sets. PLOS Computational Biology; 12(8): e1005036. https://doi.org/10.1371/journal.pcbi.1005036

Hartleb D, Fritzemeier CJ, Martin L. (2018). Automated high-quality reconstruction of metabolic networks from high-throughput data. bioRxiv: https://doi.org/10.1101/282251

Hassan SS, Williams GA, Jaiswal AK. (2019). Lignocellulosic biorefineries in europe: current state and prospects. Trends in biotechnology; 37(3): 231–234. https://doi.org/10.1016/j.tibtech.2018.07.002

Henry CS, DeJongh M, Best AA, Frybarger PM, Linsay B, Stevens RL. (2010). High-throughput generation, optimization and analysis of genome-scale metabolic models. Nature Biotechnology; 28: 977–982. https://doi.org/10.1038/nbt.1672

Ißbrücker C. (2018). How much land do we really need to produce bio-based plastics? [technical article]. European Bioplastics. 2018 Feb. 28. https://www.european-bioplastics.org/how-much-land-do-we-really-need-to-produce-bio-based-plastics/

Klamt S, Regensburger G, Gerstl MP, Jungreuthmayer C, Schuster S, Mahadevan R, et al. (2017). From elementary flux modes to elementary flux vectors: Metabolic pathway analysis with arbitrary linear flux constraints. PLOS Computational Biology; 13(4): e1005409. https://doi.org/10.1371/journal.pcbi.1005409

Klamt S, Müller S, Regensburger G, Zanghellini J. (2018). A mathematical framework for yield (vs. rate) optimization in constraint-based modeling and applications in metabolic engineering. Metabolic Engineering; 47: 153–169. https://doi.org/10.1016/j.ymben.2018.02.001

Kovárová-Kovar K, Egli T. (1998). Growth kinetics of suspended microbial cells: from single-substrate-controlled growth to mixed-substrate kinetics. Microbiology and Molecular Biology Reviews; 62(3): 646–666. https://mmbr.asm.org/content/62/3/646.full.pdf

Kumar VS, Dasika MS, Maranas CD. (2007). Optimization based automated curation of metabolic reconstructions. BMC Bioinformatics; 8: 212. https://doi.org/10.1186/1471-2105-8-212

Kurzweil R. (2004). The law of accelerating returns. In: Teuscher C, editor. Alan Turing: Life and Legacy of a Great Thinker. Heidelberg: Springer; 2004. https://doi.org/10.1007/978-3-662-05642-4_16

Latendresse M. (2014). Efficiently gap-filling reaction networks. BMC Bioinformatics; 15: 225. https://doi.org/10.1186/1471-2105-15-225

Latorre M, Cortés MP, Travisany D, Di Genova A, Budinich M, Reyes-Jara A, et al. (2016). The bioleaching potential of a bacterial consortium. Bioresource Technology 218; 659–666. https://doi.org/10.1016/j.biortech.2016.07.012

Lewis NE, Hixson KK, Conrad TM, Lerman JA, Charusanti P, Polpitiya AD, et al. (2010). Omic data from evolved *E. coli* are consistent with computed optimal growth from genome-scale models. Molecular systems biology; 6: 390. https://doi.org/10.1038/msb.2010.47

Li A, Schertzer JW, Yong X. (2018). Molecular dynamics modeling of *Pseudomonas aeruginosa* outer membranes. Physical Chemistry Chemical Physics; 20(36): 23635–23648. https://doi.org/10.1039/C8CP04278K

Lindahl, L-A. (2015). Convexity and optimization. Uppsala, Uppsala Universitet; 2015. Available from: http://www2.math.uu.se/~lal/kompendier/Convexity2015.pdf

Loira N. (2012). Scaffold-based reconstruction method for genome-scale metabolic models. [Ph. D. thesis (Bioinformatics)]. Bordeaux: Université Sciences et Technologies - Bordeaux I; 2012. Available from: https://tel.archives-ouvertes.fr/tel-00678991/document

Lv L, Ren Y-L, Chen J-C, Wu Q, Chen G-Q. (2015). Application of CRISPRi for prokaryotic metabolic engineering involving multiple genes, a case study: Controllable P(3HB-co-4HB) biosynthesis. Metabolic Engineering; 29: 160–168. https://doi.org/10.1016/j.ymben.2015.03.013

Machado D, Andrejev S, Tramontano M, Patil KR. (2018). Fast automated reconstruction of genome-scale metabolic models for microbial species and communities. Nucleic acids research; 46(15): 7542–7553. https://doi.org/10.1093/nar/gky537

Mahadevan R, Edwards JS, Doyle FJ. (2002). Dynamic flux balance analysis of diauxic growth in *Escherichia coli*. Biophysical journal; 83(3): 1331–1340. https://doi.org/10.1016/S0006-3495(02)73903-9

Mellbye BL, Giguere AT, Murthy GS, Bottomley PJ, Sayavedra-Soto LA, Chaplen FWR. (2018). Genome-scale, constraint-based modeling of nitrogen oxide fluxes during coculture of *Nitrosomonas europaea* and *Nitrobacter winogradskyi*. mSystems; 3: e00170–17. https://doi.org/10.1128/mSystems.00170-17

Milreu PV, Klein CC, Cottret L, Acuña V, Birmelé E, Borassi M, et al. (2014). Telling metabolic stories to explore metabolomics data: a case study on the yeast response to cadmium exposure. Bioinformatics; 30(1): 61–70. https://doi.org/10.1093/bioinformatics/btt597

Monod J. (1949). The growth of bacterial cultures. Annual Review of Microbiology; 3: 371–394. https://doi.org/10.1146/annurev.mi.03.100149.002103

Montano-Herrera L, Laycock B, Werker A, Pratt S. (2017). The evolution of polymer composition during PHA accumulation: the significance of reducing equivalents. Bioengineering; 4(1): 20. https://doi.org/10.3390/bioengineering4010020

Mozejko-Ciesielska J, Kiewisz R. (2016). Bacterial polyhydroxyalkanoates: still fabulous?. Microbiological Research; 192: 271–282. https://doi.org/10.1016/j.micres.2016.07.010

Muir P, Li S, Lou S, Wang D, Spakowicz DJ, Salichos L, et al. (2016). The real cost of sequencing: scaling computation to keep pace with data generation. Genome Biology; 17(1): 53. https://doi.org/10.1186/s13059-016-0917-0

Müller AC, Bockmayr A. (2013). Fast thermodynamically constrained flux variability analysis. Bioinformatics; 29(7): 903–909. https://doi.org/10.1093/bioinformatics/btt059

Mursula AM, Van Aalten DM, Hiltunen JK, Wierenga RK. (2001). The crystal structure of delta(3)-delta(2)-enoyl-CoA isomerase. Journal of Molecular Biology; 309(4): 845–53. https://doi.org/10.1006/jmbi.2001.4671

National Human Genome Research Institute: DNA sequencing costs. (2018). Available from: https://www.genome.gov/sequencingcostsdata/

Neidhardt FC, Ingraham JL, Schaechter M. (1990). Physiology of the bacterial cell: a molecular approach. Sunderland: Sinauer Associates; 1990

Nichio B, Marchaukoski JN, Raittz RT. (2017). New tools in orthology analysis: a brief review of promising perspectives. Frontiers in genetics; 8: 165. https://doi.org/10.3389/fgene.2017.00165

Nijland GO, Schouls J, Goudriaan J. (2008). Integrating the production functions of Liebig, Michaelis-Menten, Mitscherlich and Liebscher into one system dynamics model. NJAS - Wageningen Journal of Life Sciences; 55: 199–224. http://doi.org/10.1016/S1573-5214(08)80037-1

Nikel PI, Chavarría M, Fuhrer T, Sauer U, De Lorenzo V. (2015). *Pseudomonas putida* KT2440 strain metabolizes glucose through a cycle formed by enzymes of the Entner-Doudoroff, Embden-Meyerhof-Parnas, and Pentose Phosphate pathways. The Journal of biological chemistry; 290(43): 25920–25932. https://doi.org/10.1074/jbc.M115.687749

Numata K, Abe H, Iwata T. (2009). Biodegradability of poly(hydroxyalkanoate) materials. MDPI Materials; 2009(2): 1104–1126. https://doi.org/10.3390/ma2031104

Olavarria K, Marone MP, Da Costa Oliveira H, Roncallo JC, Da Costa Vasconcelos FN, Da Silva LF, et al. (2015). Quantifying NAD(P)H production in the upper Entner-Doudoroff pathway from *Pseudomonas putida* KT2440. FEBS open bio; 5: 908–915. https://doi.org/10.1016/j.fob.2015.11.002

Oliveira-Filho ER, Guamán LP, Mendonça TT, Long PF, Taciro MK, Gomez JGC, et al. (2018). Production of polyhydroxyalkanoates copolymers by recombinant *Pseudomonas* in plasmid- and antibiotic-free cultures. Journal of Molecular Microbiology and Biotechnology; 28(5): 225–235. doi: https://doi.org/10.1159/000495752

Organization for Economic Co-operation and Development (2013). Policies for bioplastics in the context of a bioeconomy. Technology and Industry Policy Papers, No. 10. Paris: OECD Publishing; 2010. https://doi.org/10.1787/5k3xpf9rrw6d-en

Orth JD, Palsson BØ. (2012). Gap-filling analysis of the iJO1366 *Escherichia coli* metabolic network reconstruction for discovery of metabolic functions. BMC systems biology; 6: 30. https://doi.org/10.1186/1752-0509-6-30

Orth JD, Thiele I, Palsson BØ. (2010). What is flux balance analysis?. Nature Biotechnology; 28(3): 245–248. https://doi.org/10.1038/nbt.1614

Parks HR. (2013). The volume of the unit n-ball. Mathematics Magazine; 86(4): 270–274. https://10.4169/math.mag.86.4.270

Pirt SJ. (1965). The maintenance energy of bacteria in growing cultures. Proceedings of the Royal Society of London; 163 (991, part B – biological sciences): 224–231. https://doi.org/10.1098/rspb.1965.0069

Pirt SJ. (1982). Maintenance energy: a general model for energy-limited and energy-sufficient growth. Archives of Microbiology; 133(4): 300–302. https://doi.org/10.1007/bf00521294

Pirt SJ. (1987). The energetics of microbes at slow growth rates: maintenance energies and dormant organisms. Journal of Fermentation Technology; 65(2): 173–177. https://doi.org/10.1016/0385-6380(87)90161-0

Poblete-Castro I, Binger D, Rodrigues A, Becker J, Martins dos Santos VAP, Wittmann C. (2013). *In-silico*-driven metabolic engineering of *Pseudomonas putida* for enhanced production of poly-hydroxyalkanoates. Metabolic Engineering; 15: 113–123. https://doi.org/10.1016/j.ymben.2012.10.004

Poblete-Castro I, Rodriguez AL, Lam CMC, Kessler W. (2014). Improved production of medium-chain-length polyhydroxyalkanoates in glucose-based fed-batch cultivations of metabolically engineered *Pseudomonas putida* strains. Journal of Microbiology and Biotechnology; 24(1): 59–69. https://doi.org/10.4014/jmb.1308.08052

Provost A, Bastin G, Agathos SN, Schneider YJ. (2006). Metabolic design of macroscopic bioreaction models: application to Chinese hamster ovary cells. Bioprocess and biosystems engineering; 29(5–6): 349–66. https://doi.org/10.1007/s00449-006-0083-y

Puchałka J, Oberhardt MA, Godinho M, Bielecka A, Regenhardt D, Timmis KN, et al (2008). Genome-scale reconstruction and analysis of the *Pseudomonas putida* KT2440 metabolic network facilitates applications in biotechnology. PLoS computational biology; 4(10): e1000210. https://doi.org/10.1371/journal.pcbi.1000210

Rancati G, Moffat J, Typas A, Pavelka N. (2017). Emerging and evolving concepts in gene essentiality. Nature Reviews Genetics; 19: 34–49. https://doi.org/10.1038/nrg.2017.74

Reimers A-M, Reimers AC. (2016). The steady-state assumption in oscillating and growing systems. Journal of Theoretical Biology; 406: 176–186. https://doi.org/10.1016/j.jtbi.2016.06.031

Rehm BH. (2003). Polyester synthases: natural catalysts for plastics. The Biochemical journal; 376(Pt 1): 15–33. https://doi.org/10.1042/BJ20031254

Riascos CAM, Gombert AK, Silva LF, Taciro MK, Gomez JGC, Le Roux GAC. (2013). Metabolic pathways analysis in PHAs production by Pseudomonas with 13 C-labeling experiments. In: Kraslawski A, Turunen I, editors. 23rd European Symposium on Computer Aided Process Engineering. Lappeenranta: Elsevier, 2013. 32 vol. p. 121–126

Ronzon T, Santini F, M'Barek R. (2015). The bioeconomy in the European Union in numbers - Facts and figures on biomass, turnover and employment. [Scientific and technical research report]. Sevilla, Joint Research Centre; 2015. https://ec.europa.eu/jrc/sites/jrcsh/files/JRC97789%20Factsheet_Bioeconomy_final.pdf

Sánchez RJ, Schripsema J, da Silva LF, Taciro MK, Pradella JG, Gomez JGC. (2003). Medium-chain-length polyhydroxyalkanoic acids (PHAmcl) produced by *Pseudomonas putida* IPT 046 from renewable sources. European Polymer Journal; 39: 1385–1394. https://doi.org/10.1016/S0014-3057(03)00019-3

Santos CI, Silvia CC, Mussatto, SI, Osseweijer P, van der Wielen LAM, Posada JA. (2018). Integrated 1st and 2nd generation sugarcane bio-refinery for jet fuel production in Brazil: Techno-economic and greenhouse gas emissions assessment. Renewable Energy; 129(part B): 733–747. https://doi.org/10.1016/j.renene.2017.05.011

Sashiwa H, Fukuda R, Okura T, Sato S, Nakayama A. (2018). Microbial degradation behavior in seawater of polyester blends containing poly(3-hydroxybutyrate-co-3-hydroxyhexanoate) (PHBHHx). Marine drugs; 16(1): 34. https://doi.org/10.3390/md16010034

Schuster S, Dandekar T, Fell DA. (1999). Detection of elementary flux modes in biochemical networks: a promising tool for pathway analysis and metabolic engineering. Trends in Biotechnology; 17(2): 53–60. https://doi.org/10.1016/S0167-7799(98)01290-6

Shrivastav A, Mishra SK, Pancha I, Jain D, Bhattacharya S, Patel S, et al. (2011). Biodegradability studies of polyhydroxyalkanoate (PHA) film produced by a marine bacteria using Jatropha biodiesel byproduct as a substrate. World Journal of Microbiology and Biotechnology; 27(7): 1531–1541. https://doi.org/10.1007/s11274-010-0605-2

Shrivastav A, Kim H-Y, Kim Y-R. (2013). Advances in the applications of polyhydroxyalkanoate nanoparticles for novel drug delivery system. BioMed Research International; 2013: 581684. https://doi.org/10.1155/2013/581684

Silva-Queiroz SR, Silva LF, Pradella JGC, Pereira EM, Gomez JGC. (2009). PHA(MCL) biosynthesis systems in *Pseudomonas aeruginosa* and *Pseudomonas putida* strains show differences on monomer specificities. Journal of Biotechnology; 143(2): 111–118. https://doi.org/10.1016/j.jbiotec.2009.06.014

Singh M, Kumar P, Ray S, Kalia VC. (2015). Challenges and opportunities for customizing polyhydroxyalkanoates. Indian journal of microbiology; 55(3): 235–249. https://doi.org/10.1007/s12088-015-0528-6

Song H, Morgan JA, Ramkrishna D. (2009). Systematic development of hybrid cybernetic models: Application to recombinant yeast co-consuming glucose and xylose. Biotechnology and Bioengineering; 103: 984–1002. https://doi.org/10.1002/bit.22332

Song H, Ramkrishna D, Pinchuk GE, Beliaev AS, Konopka AE, Fredrickson JK. (2013). Dynamic modeling of aerobic growth of *Shewanella oneidensis* – predicting triauxic growth, flux distributions, and energy requirement for growth. Metabolic Engineering; 15: 25–33. https://doi.org/10.1016/j.ymben.2012.08.004

Sudarsan S, Dethlefsen S, Blank LM, Siemann-Herzberg M, Schmid A. (2014). The functional structure of central carbon metabolism in *Pseudomonas putida* KT2440. Applied and environmental microbiology; 80(17): 5292–5303. https://doi.org/10.1128/AEM.01643-14

Sudesh K, Abe H, Doi Y. (2000). Synthesis, structure and properties of polyhydroxyalkanoates: Biological polyesters. in Progress in Polymer Science; 25(10): 1503–1555. https://doi.org/10.1016/S0079-6700(00)00035-6

Szenk M, Dill KA, De Graff AMR. (2017). Why do fast-growing bacteria enter overflow metabolism? Testing the membrane real estate hypothesis. Cell Systems; 5(2): 95–104. https://doi.org/10.1016/j.cels.2017.06.005

Taciro MK. (2008). Processo contínuo de produção de polihidroxialcanoatos de cadeia média (PHAmcl) sob limitaçao múltipla de nutrientes. [Ph. D. thesis (Biotechnology)]. São Paulo: Instituto de Ciências Biomédicas, Universidade de São Paulo; 2008. Available from: http://www.teses.usp.br/teses/disponiveis/87/87131/tde-14012009-091928/pt-br.php

Tamura T. (2018). Grid-based computational methods for the design of constraint-based parsimonious chemical reaction networks to simulate metabolite production: GridProd. BMC Bioinformatics; 19(1): 325. https://doi.org/10.1186/s12859-018-2352-6

Taymaz-Nikerel H, Borujeni AE, Verheijen PJ, Heijnen JJ., Van Gulik WM. (2010). Genome-derived minimal metabolic models for *Escherichia coli* MG1655 with estimated *in vivo* respiratory ATP stoichiometry. Biotechnology and Bioengineering; 107(2): 369–381. https://doi.org/10.1002/bit.22802

Terzer M, Maynard ND, Covert MW, Stelling J. (2009). Genome-scale metabolic networks. WIREs Systems Biology and Medicine, 1: 285–297. https://doi.org/10.1002/wsbm.37

Tsuge T. (2016). Fundamental factors determining the molecular weight of polyhydroxyalkanoate during biosynthesis. Polymer Journal; 48: 1051–1057. https://doi.org/10.1038/pj.2016.78

União da Indústria de Cana-de-Açúcar, Centro Nacional das Indústrias do Setor Sucroenergético e Biocombustíveis. (2016). Setor sucroenergético no Brasil - uma visão para 2030. São Paulo, 2016. http://www.mme.gov.br/documents/10584/7948692/UNICA-CEISE_Setor+Sucroenerg%C3%A9tico+no+Brasil_Uma+Vis%C3%A3o+para+2030.pdf

Van Bodegom P. (2007). Microbial maintenance: a critical review on its quantification. Microbial ecology; 53(4): 513–23. https://doi.org/10.1007/s00248-006-9049-5

Van Dijk EL, Jaszczyszyn Y, Naquin D, Thermes C. (2018). The third revolution in sequencing technology. Trends in Genetics; 34(9): 666–681. https://doi.org/10.1016/j.tig.2018.05.008

Van Duuren JB, Puchałka J, Mars AE, Bücker R, Eggink G, Wittmann C, et al. (2013). Reconciling *in vivo* and *in silico* key biological parameters of *Pseudomonas putida* KT2440 during growth on glucose under carbon-limited condition. BMC biotechnology; 13: 93. https://doi.org/10.1186/1472-6750-13-93

Van Heerden JH, Wortel MT, Bruggeman FJ, Heijnen JJ, Bollen YJM, Planqué R, et al. (2014). Lost in transition: start-up of glycolysis yields subpopulations of nongrowing cells. Science; 343(6174): 1245114. https://doi.org/10.1126/science.1245114

Wagner C, Urbanczik R. (2005). The geometry of the flux cone of a metabolic network. Biophysical journal; 89(6): 3837–3845. https://doi.org/10.1529/biophysj.104.055129

Wallner B, Fang H, Ohlson T, Frey-Skött J, Elofsson A. (2004). Using evolutionary information for the query and target improves fold recognition. Proteins; 54: 342–350. https://doi.org/10.1002/prot.10565

Wang G, Post WM. (2012). A theoretical reassessment of microbial maintenance and implications for microbial ecology modeling. FEMS Microbiology Ecology; 81(3): 610–617. https://doi.org/10.1111/j.1574-6941.2012.01389.x

Wiggins S, Keats S, Compton J. (2010). What caused the food price spike of 2007/08? Lessons for world cereals markets. London, Overseas Development Institute; 2010. https://www.odi.org/sites/odi.org.uk/files/odi-assets/publications-opinion-files/6103.pdf

Yousofshahi M, Ullah E, Stern R, Hassoun S. (2013). MC3: a steady-state model and constraint consistency checker for biochemical networks. BMC systems biology; 7: 129. https://doi.org/10.1186/1752-0509-7-129

Yuan Q, Huang T, Li P, Hao T, Li F, Ma H, et al. (2017). Pathway-consensus approach to metabolic network reconstruction for *Pseudomonas putida* KT2440 by systematic comparison of published models. PLOS ONE; 12(1): e0169437. https://doi.org/10.1371/journal.pone.0169437

Zanghellini J, Ruckerbauer DE, Hanscho M, Jungreuthmayer C. (2013). Elementary flux modes in a nutshell: properties, calculation and applications. Biotechnology Journal; 8: 1009–1016. https://doi.org/10.1002/biot.201200269

Zinn M, Witholt B, Egli T. (2004). Dual nutrient limited growth: models, experimental observations and applications. Journal of Biotechnology; 113(1–3): 263–279. https://doi.org/10.1016/j.jbiotec.2004.03.030

Ziv N, Brandt NJ, Gresham D. (2013). The use of chemostats in microbial systems biology. Journal of visualized experiments (JoVE); 80: 50168. https://doi.org/10.3791/50168

# APPENDIX A - Log of the manual method

The draft genome-scale metabolic network of *Pseudomonas* sp. LFM046 was reconstructed automatically from the genome in the KBase platform. To unblock the biomass reaction and start the manual refinement method, the most basic gapfilling method from KBase was applied: the "complete media", which allows influxes from any exchange reactions.

Then, 12 iterations of the manual method were performed. An iteration is represented here as a table containing the exchange set (set of exchange reactions with non-null flux), the wrong influx chosen to be resolved and the modifications in the network connected to that reaction.

<table>
<tr><th colspan="5">Iteration 01</th></tr>
<tr><th colspan="5">Exchange set (chosen wrong influx in grey)</th></tr>
<tr><td>H2O</td><td>L-Lysine</td><td>D-Fructose</td><td>Aminoethanol</td><td rowspan="2">Nicotinamide ribonucleotide</td></tr>
<tr><td>O2</td><td>L-Aspartate</td><td>Cl-</td><td>K+</td></tr>
<tr><td>Phosphate</td><td>Sulfate</td><td>L-Leucine</td><td>Riboflavin</td><td>Folate</td></tr>
<tr><td>CO2</td><td>L-Arginine</td><td>Putrescine</td><td>HYXN</td><td>TRHL</td></tr>
<tr><td>NH3</td><td>Cu2+</td><td>L-Histidine</td><td>Mg</td><td>ocdca</td></tr>
<tr><td>L-Glutamate</td><td>L-Methionine</td><td>L-Proline</td><td>Spermidine</td><td>Myristic acid</td></tr>
<tr><td>2-Oxoglutarate</td><td>Ca2+</td><td>L-Malate</td><td>H2S2O3</td><td>fe3</td></tr>
<tr><td>D-Glucose</td><td>L-Tryptophan</td><td>D-Mannose</td><td>GLUM</td><td>Biomass</td></tr>
<tr><td>Mn2+</td><td>L-Phenylalanine</td><td>Acetoacetate</td><td>Thiamin</td><td rowspan="2">Glycerol-3-phosphate</td></tr>
<tr><td>Glycine</td><td>H+</td><td>Co2+</td><td>Cytosine</td></tr>
<tr><td>Zn2+</td><td>L-Tyrosine</td><td>L-Valine</td><td>XAN</td><td></td></tr>
<tr><td>Succinate</td><td>Menaquinone 7</td><td>L-Lactate</td><td>L-Isoleucine</td><td></td></tr>
<tr><th colspan="5">Modifications in the network</th></tr>
<tr><td colspan="5">1. Add reactions to produce N-Succinyl-L-2,6-diaminopimelate (cpd02698)<br>- info: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3862034/<br>- using the succinylase pathway (with TPN = NADP)<br>- added reactions (SEED ID's): rxn01644, rxn02928, rxn03087</td></tr>
</table>

| Iteration 02 | | | | |
|---|---|---|---|---|
| Exchange set (chosen wrong influx in grey) | | | | |
| H2O | Sulfate | Uracil | Mg | Myristic acid |
| O2 | Cu2+ | Cl- | Spermidine | fe3 |
| Phosphate | Ca2+ | Fumarate | Thiamin | Biomass |
| CO2 | Ornithine | Co2+ | Cytosine | ala-L-Thr-L |
| NH3 | H+ | K+ | Folate | Ala-His |
| D-Glucose | Urea | Nitrate | ocdca | Menaquinone 7 |
| Mn2+ | Nitrite | Butyrate | L-Methionine S-oxide | Gly-Leu |
| Zn2+ | D-Fructose | Riboflavin | | Gly-Phe |
| Modifications in the network | | | | |
| 1. Complete the Thiamine (vitamin B1) synthesis pathway. Using this pathway because its final reactions were found in the GPR's of the genome: http://www.pseudomonas.com:1555/PSEUDO/NEW-IMAGE?object=THISYN-PWY<br>- added reactions: rxn08131, rxn10095, rxn01538, rxn12239, rxn00958, rxn00957, rxn00960, rxn00962, rxn02143, rxn00598 | | | | |

| Iteration 03 | | | | |
|---|---|---|---|---|
| Exchange set (chosen wrong influx in grey) | | | | |
| H2O | Sulfate | Uracil | Mg | ala-L-Thr-L |
| O2 | Cu2+ | Cl- | Spermidine | Ala-His |
| Phosphate | Ca2+ | Fumarate | Cytosine | Gly-Met |
| CO2 | Ornithine | Co2+ | Folate | Menaquinone 7 |
| NH3 | H+ | K+ | ocdca | Gly-Leu |
| D-Glucose | Urea | Nitrate | Myristic acid | Gly-Phe |
| Mn2+ | Nitrite | Butyrate | fe3 | Gly-Tyr |
| Zn2+ | D-Fructose | Riboflavin | Biomass | |
| Modifications in the network | | | | |
| 1. Add a reaction to produce cytosine<br>- info: https://en.wikipedia.org/wiki/Biosynthesis<br>- added reactions: rxn00362 | | | | |

Iteration 04

Exchange set (chosen wrong influx in grey)

| | | | | |
|---|---|---|---|---|
| H2O | Zn2+ | D-Fructose | Spermidine | Ala-Leu |
| O2 | Sulfate | Uracil | Folate | Menaquinone 7 |
| Phosphate | Cu2+ | Cl- | ocdca | Gly-Phe |
| CO2 | Ca2+ | Co2+ | L-Methionine S-oxide | Myristic acid |
| NH3 | Ornithine | K+ | | |
| D-Glucose | H+ | Riboflavin | fe3 | |
| Mn2+ | Urea | Mg | Biomass | |

Modifications in the network

1. Add a reaction to produce ornithine because it was blocked and that seemed to be blocking also the proline synthesis
- info: https://en.wikipedia.org/wiki/Biosynthesis
- added reactions: rxn01637

Iteration 05

Exchange set (chosen wrong influx in grey)

| | | | | |
|---|---|---|---|---|
| H2O | Zn2+ | Co2+ | L-Methionine S-oxide | Menaquinone 7 |
| O2 | Sulfate | K+ | | Gly-Phe |
| Phosphate | Cu2+ | Riboflavin | Myristic acid | Gly-Gln |
| CO2 | Ca2+ | Mg | fe3 | |
| NH3 | H+ | Spermidine | Biomass | |
| D-Glucose | D-Fructose | Folate | gly-asn-L | |
| Mn2+ | Cl- | ocdca | Ala-Leu | |

Modifications in the network

1. Fatty-acids *de novo* synthesis had to be rebuilt due to several problems:
- no production of trans-octodec-2-enoyl-ACP
- all the reactions that convert R-3-hydroxy to enoyl (same enzyme for all)
- palmitoyl and hexadecanoyl are the same thing but had different ID's
- 4 compounds with 18 carbons had two versions each: "octo" and "octa", 3 of them also had false duplicates (messed up ACP and organic radical formulas)
- reactions rxn05322, rxn05323, rxn05324 and rxn05327 had wrong stoichiometry and their ACP compounds had charge -2 instead of -1 like they have in SEED database
- added reactions: rxn08396, rxn08397, rxn05462, rxn08392, rxn08393, rxn05332, rxn05351, rxn05335, rxn05331, rxn05333, rxn05334, rxn05330, rxn05329, rxn05355, rxn05356, rxn05352, rxn05357, rxn05353

Iteration 06

Exchange set (chosen wrong influx in grey)

| | | | | |
|---|---|---|---|---|
| H+ | NH3 | fe3 | Ala-Leu | Cl- |
| D-Fructose | Phosphate | Mg | D-Glucose | Mn2+ |
| H2O | Zn2+ | K+ | CO2 | Folate |
| Ca2+ | Gly-Phe | Sulfate | Biomass | Riboflavin |
| Cu2+ | Spermidine | Co2+ | Menaquinone 7 | |

Modifications in the network

1. $O_2$ was being excreted instead of consumed.
2. There were no obvious reasons, so the dipeptides and folate were chosen. The dipeptides were being broken to supply aminoacids:

- glycine was the only complete pathway, the other 3 aminoacids and folate all had 1 reaction missing each
- Alanine: rxn00191
- Folate: rxn03167, info: https://www.wikipathways.org/index.php/Pathway:WP2360
- Leucine: rxn01208
- Phenylalanine: rxn01256

Iterations 07 and 08

Exchange set (chosen wrong influx in grey)

| | | | | |
|---|---|---|---|---|
| H2O | D-Glucose | Ca2+ | Co2+ | fe3 |
| O2 | Mn2+ | H+ | K+ | Biomass |
| Phosphate | Zn2+ | Urea | Riboflavin | Menaquinone 7 |
| CO2 | Sulfate | D-Fructose | Mg | |
| NH3 | Cu2+ | Cl- | Spermidine | |

Modifications in the network

1. $O_2$ is not wrong because of its value but because it had to be forced to have an influx value (in other words, left by itself it had zero flux). One plausible cause was a high influx of $Fe^{3+}$ being reduced to $Fe^{2+}$ and oxidizing $H_2O$ to $O_2$, providing $O_2$ to the metabolism

- added reactions: rxn01453, rxn12757, rxn 12758, rxn12752, rxn12753, rxn09280, rxn07997, rxn07987, rxn12754, R_A_PHA_W, R_A_DNA_W
- the last two reactions are artificial (PHA and DNA composition estimated for LFM046). The others are from PHA synthesis

Iteration 09

Exchange set (chosen wrong influx in grey)

| | | | | |
|---|---|---|---|---|
| H2O | D-Glucose | Ca2+ | Co2+ | fe3 |
| O2 | Mn2+ | H+ | K+ | Biomass |
| Phosphate | Zn2+ | Urea | Riboflavin | Menaquinone 7 |
| CO2 | Sulfate | D-Fructose | Mg | |
| NH3 | Cu2+ | Cl- | Spermidine | |

Modifications in the network

1. $O_2$ is not wrong because of its value but because it had to be forced to have an influx value (in other words, left by itself it had zero flux). One plausible cause was a high influx of $Fe^{3+}$ being reduced to $Fe^{2+}$ and oxidizing $H_2O$ to $O_2$, providing $O_2$ to the metabolism

- deleted reaction added in iteration 07 (rxn09295) because it was found the model already had an equivalent reaction (rxn10125)
- added ATP sink for traditional maintenance model: R_A_MAIN

Iteration 10

Exchange set (omitting $H^+$)

| | | | | |
|---|---|---|---|---|
| D-Fructose | NH3 | O2 | H2O | Sulfate |
| D-Glucose | Phosphate | CO2 | Biomass | fe3 |

Modifications in the network

1. $O_2$ was not wrong because of its value but because it had to be forced to have an influx value (in other words, left by itself it had zero flux). One plausible cause was a high influx of $Fe^{3+}$ being reduced to $Fe^{2+}$ and oxidizing $H_2O$ to $O_2$, providing $O_2$ to the metabolism. Tested changing the lower and upper bounds of the exchange reactions of glucose, fructose, $O_2$, $CO_2$, $NH_4^+$ and $PO_4^{3-}$ as well as of the R_A_MAIN to values close to those of an experiment

2. Excluded from the exchange set reactions with a very low flux (like vitamins and ions). H+ is in the exchange set but is omitted from this iteration for simplification

With iteration 10, it was found that there were no more wrong influxes and the problem of $O_2$ was probably caused by too large bounds of exchange reactions like glucose and $NH_4^+$ in relationship to the exchange reactions of $H^+$ and $H_2O$, making the latter limiting for the biomass maximization. For example, if $H_2O$ is being produced and its upper bound is limiting the maximization, a set of reactions whose net effect is to oxidize $H_2O$ into $O_2$ will be activated.

This example is an Energy-Generating Cycle (EGC) because this net effect would require supply of external energy *in vivo*. However, unlike a cycle that produces ATP from ADP, this one is not as explicit, even though the root cause is the same: reactions with wrong directions. Also, in this case the cycle is avoided simply by setting safe lower and upper bounds of the exchange reactions, what is not the general case.

The iterations 10b to 12e are simply tests of changes in the stoichiometry of the biomass reaction and setting the bounds of the exchange reactions with measured fluxes to values of different experimental points. In the network structure, the only change was the addition of reactions that metabolize glucose and fructose via conversion to gluconate, to test their effect in the flux distribution (how much of the carbon enters the cell via gluconate). These reactions, which do not change the prediction of external fluxes, are: rxn10116, rxn13790, rxn12740, rxn05571, rxn12783, rxn01921 and rxn01474.

To evaluate the quality of the predictions of the resulting network, four estimators were calculated using the 7 measured fluxes of each point: average relative error (ARE), average magnitude of the relative error (AMRE), maximum magnitude of the relative error or maximum predictive error (MPE) and average euclidean distance per coordinate (AEDC).

The 7 measured fluxes (coordinates) are: biomass (including PHA), fructose, glucose, $NH_4^+$, $PO_4^{3-}$, $O_2$ and $CO_2$. The polyhydroxyalkanoate (PHA) content was included directly in the biomass export equation. For each flux, the relative error is 0 if the measured value is 0 or simulated/measured - 1 otherwise. ARE is the simple average of these errors, AMRE is the average of their absolute values, MPE is the maximum of the absolute values and AEDC is the magnitude of the vector formed by them divided by 7. The results are in Table A.1.

Table A.1: Measured flux values and prediction error estimators after 12 iterations

| Flux \ version | 12a | 12b | 12c | 12d | 12e |
|---|---|---|---|---|---|
| Steady-state ID | 1101 | 1201 | 701 | 304 | 1401 |

| | | | | | |
|---|---|---|---|---|---|
| Lim. nutrients | {N} | {P} | {P} | {N; P} | {C; N; P} |
| No. of samples | 6 | 7 | 10 | 6 | 6 |
| Biomass | +0.098 | +0.063 | +0.049 | +0.050 | 0.512 |
| Fructose | 0 | 0 | -1.746 | -0.508 | -2.155 |
| Glucose | -2.219 | -2.197 | -1.576 | -2.031 | -4.039 |
| $NH_4^+$ | -0.878 | -1.150 | -0.653 | -0.606 | -5.655 |
| $PO_4^{3-}$ | -0.100 | -0.038 | -0.052 | -0.098 | -0.390 |
| $O_2$ | -6.851 | -5.309 | -10.59 | -5.110 | -13.02 |
| $CO_2$ | +7.541 | +5.591 | +10.87 | +7.122 | +11.98 |
| | | | Error estimators (%) | | |
| C mass error | -5.6 ± 3.5 | +22 ± 1.9 | +5.6 ± 1.9 | +9.2 ± 2.1 | +16 ± 1.8 |
| ARE | -3.43 | -2.19 | -3.04 | +1.36 | +0.46 |
| AMRE | 4.38 | 4.86 | 3.73 | 1.50 | 1.67 |
| MPE | 9.29 | 22.1 | 18.0 | 10.0 | 2.47 |
| AEDC | 2.22 | 3.29 | 2.64 | 0.63 | 0.68 |

AEDC is the only multivariate prediction error estimator among the 4 defined here. Thus it is the most adequate to evaluate the global quality of the phenotypic predictions of a metabolic network. As shown in table 1, the experiments 304 and 1401 are the best predicted ones among the 5 tested, despite not being the most reliable ones according to the independent carbon mass balance error. They are also the only ones with multiple-limitation regimes.

This may be because the vast majority of the Gene-Protein-Reaction associations (GPR) as well as the biomass composition data in the literature are validated with experiments of carbon limitation or no limitation at all (equivalent to equal limitation of all non-interchangeable nutrients). This result suggests that most if not all genome-scale reconstructions today are biased towards the condition of carbon (energy) or multiple limitation since their draft states.

If this is true, the *N*-phenotypic model and its application in metabolic network refining (such as in *N*-GlobalFit) is one way to detect and correct this bias, not only in individual models but also in the public databases that are used to build and refine them. Yet another evidence of this bias is the usual assumption that any biomass can be modelled with the composition of *E. coli*, what hides the underlying assumption that small changes in the biomass formation stoichiometry do not significantly affect the predictive quality of a model:

- The underlying assumption may hold for cultures under carbon limitation, because carbon makes up for about half of the biomass and non-limiting nutrients that are not interchangeable with carbon do not affect the biomass production;
- The more concomitant non-interchangeable nutrients are limiting, the more restricted the metabolism is. And also the more similar to the scenario of energy limitation, because energy is the most limiting factor due to the maintenance process. Thus, multiple-limitation is similar to the carbon-limitation for chemoorganoheterotrophs, which are most of the organisms in literature. Thus, the assumption may also hold for multiple-limitation or no limitation at all, which really is energy limitation;
- The case-study of *Pseudomonas* sp. LFM046 with phenotypic data of single non-carbon limitation is a counterexample that violates the assumption, because small changes in the biomass content of phosphorus and nitrogen dramatically change their influxes for the same biomass outlfux.

# APPENDIX B - F.A.M.E. FBA commands

The FBA and FVA simulations were carried-out in the webservice F.A.M.E, available from: http://f-a-m-e.fame-vu.surf-hosted.nl/ajax/page1.php. Not only it does not require any kind of installation as it has its own very simple scripting language. The commands are self explanatory but are explained in the manual available at the same website (version from 25/01/2014, section 3.6). This is convenient because the list of commands is exactly reproducible and serves as a log of the metabolic network editing and simulation procedures.

## B.1 Constrain the Energy-Generating Cycles (EGC's)

```
SETCONSTRAINTS R_rxn00077_c0 0 1000
SETCONSTRAINTS R_rxn00085_c0 -1000 1000
SETCONSTRAINTS R_rxn00145_c0 -1000 1000
SETCONSTRAINTS R_rxn00182_c0 -1000 1000
SETCONSTRAINTS R_rxn00184_c0 -1000 1000
SETCONSTRAINTS R_rxn00189_c0 0 1000
SETCONSTRAINTS R_rxn00191_c0 -1000 1000
SETCONSTRAINTS R_rxn00239_c0 -1000 1000
SETCONSTRAINTS R_rxn00248_c0 -1000 1000
SETCONSTRAINTS R_rxn00260_c0 -1000 1000
SETCONSTRAINTS R_rxn00278_c0 -1000 1000
SETCONSTRAINTS R_rxn00283_c0 -1000 1000
SETCONSTRAINTS R_rxn00347_c0 -1000 1000
SETCONSTRAINTS R_rxn00379_c0 -1000 1000
SETCONSTRAINTS R_rxn00441_c0 0 1000
SETCONSTRAINTS R_rxn00499_c0 -1000 1000
SETCONSTRAINTS R_rxn00500_c0 -1000 1000
SETCONSTRAINTS R_rxn00604_c0 -1000 1000
SETCONSTRAINTS R_rxn00611_c0 -1000 1000
SETCONSTRAINTS R_rxn00612_c0 -1000 1000
SETCONSTRAINTS R_rxn00615_c0 -1000 1000
SETCONSTRAINTS R_rxn00616_c0 -1000 1000
SETCONSTRAINTS R_rxn00747_c0 -1000 1000
SETCONSTRAINTS R_rxn00763_c0 -1000 1000
SETCONSTRAINTS R_rxn00786_c0 -1000 1000
SETCONSTRAINTS R_rxn00799_c0 -1000 1000
SETCONSTRAINTS R_rxn00804_c0 -1000 1000
SETCONSTRAINTS R_rxn00806_c0 -1000 1000
SETCONSTRAINTS R_rxn00839_c0 -1000 1000
SETCONSTRAINTS R_rxn00903_c0 -1000 1000
SETCONSTRAINTS R_rxn00904_c0 -1000 1000
SETCONSTRAINTS R_rxn00910_c0 -1000 1000
SETCONSTRAINTS R_rxn00929_c0 -1000 1000
SETCONSTRAINTS R_rxn00935_c0 -1000 1000
SETCONSTRAINTS R_rxn00973_c0 -1000 1000
SETCONSTRAINTS R_rxn00974_c0 -1000 1000
SETCONSTRAINTS R_rxn01241_c0 -1000 1000
SETCONSTRAINTS R_rxn01280_c0 -1000 1000
SETCONSTRAINTS R_rxn01281_c0 -1000 1000
SETCONSTRAINTS R_rxn01301_c0 -1000 1000
SETCONSTRAINTS R_rxn01302_c0 -1000 1000
SETCONSTRAINTS R_rxn01313_c0 -1000 1000
SETCONSTRAINTS R_rxn01314_c0 -1000 1000
SETCONSTRAINTS R_rxn01388_c0 -1000 1000
SETCONSTRAINTS R_rxn01451_c0 -1000 1000
SETCONSTRAINTS R_rxn01452_c0 -1000 1000
SETCONSTRAINTS R_rxn01492_c0 -1000 1000
SETCONSTRAINTS R_rxn01512_c0 -1000 1000
SETCONSTRAINTS R_rxn01573_c0 -1000 1000
SETCONSTRAINTS R_rxn01575_c0 -1000 1000
SETCONSTRAINTS R_rxn01578_c0 -1000 1000
SETCONSTRAINTS R_rxn01579_c0 -1000 1000
SETCONSTRAINTS R_rxn01870_c0 -1000 1000
SETCONSTRAINTS R_rxn01872_c0 -1000 1000
SETCONSTRAINTS R_rxn01975_c0 -1000 1000
SETCONSTRAINTS R_rxn01977_c0 -1000 1000
SETCONSTRAINTS R_rxn02186_c0 -1000 1000
SETCONSTRAINTS R_rxn02187_c0 -1000 1000
SETCONSTRAINTS R_rxn02376_c0 -1000 1000
SETCONSTRAINTS R_rxn03068_c0 -1000 1000
SETCONSTRAINTS R_rxn03536_c0 -1000 1000
SETCONSTRAINTS R_rxn03990_c0 -1000 1000
SETCONSTRAINTS R_rxn03991_c0 -1000 1000
SETCONSTRAINTS R_rxn04413_c0 -1000 1000
SETCONSTRAINTS R_rxn04954_c0 -1000 1000
SETCONSTRAINTS R_rxn05145_c0 -1000 1000
SETCONSTRAINTS R_rxn05156_c0 0 1000
SETCONSTRAINTS R_rxn05206_c0 -1000 1000
SETCONSTRAINTS R_rxn05209_c0 -1000 1000
SETCONSTRAINTS R_rxn05217_c0 -1000 1000
SETCONSTRAINTS R_rxn05221_c0 -1000 1000
SETCONSTRAINTS R_rxn05297_c0 -1000 1000
SETCONSTRAINTS R_rxn05298_c0 -1000 1000
SETCONSTRAINTS R_rxn05303_c0 -1000 1000
SETCONSTRAINTS R_rxn05305_c0 -1000 1000
SETCONSTRAINTS R_rxn05313_c0 -1000 1000
SETCONSTRAINTS R_rxn05561_c0 -1000 1000
SETCONSTRAINTS R_rxn05596_c0 -1000 1000
SETCONSTRAINTS R_rxn05605_c0 -1000 1000
SETCONSTRAINTS R_rxn05654_c0 -1000 1000
SETCONSTRAINTS R_rxn08094_c0 -1000 1000
SETCONSTRAINTS R_rxn08291_c0 -1000 1000
SETCONSTRAINTS R_rxn08527_c0 -1000 1000
SETCONSTRAINTS R_rxn08783_c0 0 1000
SETCONSTRAINTS R_rxn08900_c0 0 1000
SETCONSTRAINTS R_rxn09188_c0 -1000 1000
SETCONSTRAINTS R_rxn09240_c0 -1000 1000
SETCONSTRAINTS R_rxn09272_c0 -1000 1000
SETCONSTRAINTS R_rxn09562_c0 -1000 1000
SETCONSTRAINTS R_rxn09674_c0 -1000 1000
SETCONSTRAINTS R_rxn10042_c0 -1000 1000
SETCONSTRAINTS R_rxn10060_c0 -1000 1000
SETCONSTRAINTS R_rxn10122_c0 -1000 1000
SETCONSTRAINTS R_rxn10131_c0 -1000 1000
SETCONSTRAINTS R_rxn10151_c0 -1000 1000
SETCONSTRAINTS R_rxn10152_c0 -1000 1000
SETCONSTRAINTS R_rxn10153_c0 -1000 1000
SETCONSTRAINTS R_rxn10154_c0 -1000 1000
SETCONSTRAINTS R_rxn10806_c0 0 1000
SETCONSTRAINTS R_rxn10945_c0 -1000 1000
```

## *B.2 Constrain the biomass micronutrients*

```
SETCONSTRAINTS EX_cpd00030_e0 -1000 1000
SETCONSTRAINTS EX_cpd00034_e0 -1000 1000
SETCONSTRAINTS EX_cpd00048_e0 -1000 1000
SETCONSTRAINTS EX_cpd00058_e0 -1000 1000
SETCONSTRAINTS EX_cpd00063_e0 -1000 1000
SETCONSTRAINTS EX_cpd00099_e0 -1000 1000
SETCONSTRAINTS EX_cpd00149_e0 -1000 1000
SETCONSTRAINTS EX_cpd00205_e0 -1000 1000
SETCONSTRAINTS EX_cpd00220_e0 -1000 1000
SETCONSTRAINTS EX_cpd00254_e0 -1000 1000
SETCONSTRAINTS EX_cpd00264_e0 -1000 1000
SETCONSTRAINTS EX_cpd10516_e0 -1000 1000
SETCONSTRAINTS EX_cpd11606_e0 -1000 1000

SETCONSTRAINTS EX_cpd00001_e0 -1000 1000
SETCONSTRAINTS EX_cpd00067_e0 -1000 1000
SETCONSTRAINTS EX_BIOTOT_e0 -1000 1000
```

## *B.3 Reset the biomass and PHA constraints*

```
SETCONSTRAINTS A_BIOTOT01_u 0 0
SETCONSTRAINTS A_BIOTOT02_u 0 0
SETCONSTRAINTS A_BIOTOT03_u 0 0
SETCONSTRAINTS A_BIOTOT04_u 0 0
SETCONSTRAINTS A_BIOTOT05_u 0 0
SETCONSTRAINTS A_BIOTOT06_u 0 0
SETCONSTRAINTS A_BIOTOT07_u 0 0
SETCONSTRAINTS A_BIOTOT08_u 0 0
SETCONSTRAINTS A_BIOTOT09_u 0 0
SETCONSTRAINTS A_BIOTOT10_u 0 0
SETCONSTRAINTS A_BIOTOT11_u 0 0
SETCONSTRAINTS A_BIOTOT01_d 0 0
SETCONSTRAINTS A_BIOTOT02_d 0 0
SETCONSTRAINTS A_BIOTOT03_d 0 0
SETCONSTRAINTS A_BIOTOT04_d 0 0
SETCONSTRAINTS A_BIOTOT05_d 0 0
SETCONSTRAINTS A_BIOTOT06_d 0 0
SETCONSTRAINTS A_BIOTOT07_d 0 0
SETCONSTRAINTS A_BIOTOT08_d 0 0
SETCONSTRAINTS A_BIOTOT09_d 0 0
SETCONSTRAINTS A_BIOTOT10_d 0 0
```

## *B.4 Test the primary hypothesis*

Commands from subsections B.1 to B.3 plus the following:

```
BATCHRUN
SETCONSTRAINTS EX_cpd00027_e0 -20 0
SETCONSTRAINTS EX_cpd00082_e0 -20 0
SETCONSTRAINTS EX_cpd00007_e0 -50 0
SETCONSTRAINTS EX_cpd00013_e0 -15 0
SETCONSTRAINTS EX_cpd00009_e0 -1 0
SETCONSTRAINTS EX_cpd00011_e0 50 100

DELOBJECTIVE LASTOBJECTIVE
ADDOBJECTIVE R_rxn05343_c0 maximize

SETCONSTRAINTS A_BIOTOT01_u 0 0
RUNFBA BIOTOT_0
SETCONSTRAINTS A_BIOTOT01_u 0.2 0.2
RUNFBA BIOTOT_0.2
SETCONSTRAINTS A_BIOTOT01_u 0.4 0.4
RUNFBA BIOTOT_0.4
SETCONSTRAINTS A_BIOTOT01_u 0.6 0.6
RUNFBA BIOTOT_0.6
SETCONSTRAINTS A_BIOTOT01_u 0.8 0.8
RUNFBA BIOTOT_0.8
SETCONSTRAINTS A_BIOTOT01_u 1 1
RUNFBA BIOTOT_1
SETCONSTRAINTS A_BIOTOT01_u 1.2 1.2
RUNFBA BIOTOT_1.2
SETCONSTRAINTS A_BIOTOT01_u 1.4 1.4
RUNFBA BIOTOT_1.4
SETCONSTRAINTS A_BIOTOT01_u 1.42 1.42
RUNFBA BIOTOT_1.42
SETCONSTRAINTS A_BIOTOT01_u 1.45 1.45
RUNFBA BIOTOT_1.45

SETCONSTRAINTS EX_cpd00009_e0 -2 0

SETCONSTRAINTS A_BIOTOT01_u 1.42 1.42
RUNFBA BIOTOT_1.42
SETCONSTRAINTS A_BIOTOT01_u 1.45 1.45
RUNFBA BIOTOT_1.45
SETCONSTRAINTS A_BIOTOT01_u 1.47 1.47
RUNFBA BIOTOT_1.47
```

Also, all these commands were repeated for the following two cases of the SBML model version 18:

Case 1) PHA with 95 %mol of $C_4$:

```
<reaction id="R_A_PHA_simulation1"  reversible="true">
    <notes>
        <html xmlns="http://www.w3.org/1999/xhtml"></html>
    </notes>
    <listOfReactants>
        <speciesReference species="M_MNXM233" stoichiometry="0.95"/>
        <speciesReference species="M_MNXM4929" stoichiometry="0.01"/>
        <speciesReference species="M_MNXM4930" stoichiometry="0.01"/>
        <speciesReference species="M_MNXM4928" stoichiometry="0.01"/>
        <speciesReference species="M_MNXM6020" stoichiometry="0.01"/>
        <speciesReference species="M_MNXM4926" stoichiometry="0.01"/>
    </listOfReactants>
    <listOfProducts>
        <speciesReference species="A_PHA_SIM1" stoichiometry="1"/>
        <speciesReference species="M_MNXM12" stoichiometry="1"/>
    </listOfProducts>
    <kineticLaw>
        <math xmlns="http://www.w3.org/1998/Math/MathML">
            <ci>FLUX_VALUE</ci>
        </math>
        <listOfParameters>
            <parameter id="LOWER_BOUND" value="0"/>
            <parameter id="UPPER_BOUND" value="1000"/>
            <parameter id="OBJECTIVE_COEFFICIENT" value="0"/>
            <parameter id="FLUX_VALUE" value="0"/>
        </listOfParameters>
    </kineticLaw>
</reaction>

<reaction id="R_A_PHA_SELECTOR"  reversible="true">
    <notes>
        <html xmlns="http://www.w3.org/1999/xhtml"></html>
    </notes>
    <listOfReactants>
        <speciesReference species="A_PHA_SIM1" stoichiometry="1.8385660919018"/>
    </listOfReactants>
    <listOfProducts>
        <speciesReference species="A_PHA" stoichiometry="1"/>
    </listOfProducts>
    <kineticLaw>
        <math xmlns="http://www.w3.org/1998/Math/MathML">
            <ci>FLUX_VALUE</ci>
        </math>
        <listOfParameters>
            <parameter id="LOWER_BOUND" value="0"/>
            <parameter id="UPPER_BOUND" value="1000"/>
            <parameter id="OBJECTIVE_COEFFICIENT" value="0"/>
            <parameter id="FLUX_VALUE" value="0"/>
        </listOfParameters>
    </kineticLaw>
</reaction>
```

Case 2) PHA with the composition of the wildtype LFM046 strain:

```
<reaction id="R_A_PHA_W_c0"  reversible="true">
    <notes>
        <html xmlns="http://www.w3.org/1999/xhtml"></html>
    </notes>
    <listOfReactants>
        <speciesReference species="M_MNXM4929" stoichiometry="0.029"/>
        <speciesReference species="M_MNXM4930" stoichiometry="0.234"/>
        <speciesReference species="M_MNXM4928" stoichiometry="0.609"/>
        <speciesReference species="M_MNXM6020" stoichiometry="0.091"/>
        <speciesReference species="M_MNXM4926" stoichiometry="0.037"/>
    </listOfReactants>
    <listOfProducts>
        <speciesReference species="A_PHA_LFM046" stoichiometry="1"/>
        <speciesReference species="M_MNXM12" stoichiometry="1"/>
    </listOfProducts>
    <kineticLaw>
        <math xmlns="http://www.w3.org/1998/Math/MathML">
            <ci>FLUX_VALUE</ci>
        </math>
        <listOfParameters>
            <parameter id="LOWER_BOUND" value="0"/>
            <parameter id="UPPER_BOUND" value="1000"/>
            <parameter id="OBJECTIVE_COEFFICIENT" value="0"/>
            <parameter id="FLUX_VALUE" value="0"/>
        </listOfParameters>
    </kineticLaw>
</reaction>

<reaction id="R_A_PHA_SELECTOR"  reversible="true">
    <notes>
        <html xmlns="http://www.w3.org/1999/xhtml"></html>
    </notes>
    <listOfReactants>
        <speciesReference species="A_PHA_LFM046" stoichiometry="1"/>
    </listOfReactants>
    <listOfProducts>
        <speciesReference species="A_PHA" stoichiometry="1"/>
    </listOfProducts>
    <kineticLaw>
        <math xmlns="http://www.w3.org/1998/Math/MathML">
            <ci>FLUX_VALUE</ci>
        </math>
        <listOfParameters>
            <parameter id="LOWER_BOUND" value="0"/>
            <parameter id="UPPER_BOUND" value="1000"/>
            <parameter id="OBJECTIVE_COEFFICIENT" value="0"/>
            <parameter id="FLUX_VALUE" value="0"/>
        </listOfParameters>
    </kineticLaw>
</reaction>
```

## *B.5 Test the secondary hypothesis*

Commands from subsections B.1 to B.3 plus the following:

```
BATCHRUN
SETCONSTRAINTS EX_cpd00027_e0 0 0
SETCONSTRAINTS EX_cpd00082_e0 0 0
SETCONSTRAINTS EX_cpd00007_e0 -50 0
SETCONSTRAINTS EX_cpd00013_e0 -15 0
SETCONSTRAINTS EX_cpd00009_e0 -1 0
SETCONSTRAINTS EX_cpd00011_e0 50 100

SETCONSTRAINTS A_CINTAKE01 0 0
SETCONSTRAINTS A_CINTAKE02 0 0
SETCONSTRAINTS A_CINTAKE03 0 0
SETCONSTRAINTS A_CINTAKE04 0 0
SETCONSTRAINTS A_CINTAKE05 0 0
SETCONSTRAINTS A_CINTAKE06 0 0
SETCONSTRAINTS A_CINTAKE07 0 0

DELOBJECTIVE LASTOBJECTIVE
SETCONSTRAINTS A_BIOTOT01_u 1.4 1.4
SETCONSTRAINTS R_rxn10125_c0 0 0

SETCONSTRAINTS A_CINTAKE01 -40 0
ADDOBJECTIVE A_CINTAKE01 maximize
RUNFBA CINTAKE01
DELOBJECTIVE LASTOBJECTIVE
SETCONSTRAINTS A_CINTAKE01 0 0

SETCONSTRAINTS A_CINTAKE02 -40 0
ADDOBJECTIVE A_CINTAKE02 maximize
RUNFBA CINTAKE02
DELOBJECTIVE LASTOBJECTIVE
SETCONSTRAINTS A_CINTAKE02 0 0

SETCONSTRAINTS A_CINTAKE03 -40 0
ADDOBJECTIVE A_CINTAKE03 maximize
RUNFBA CINTAKE03
DELOBJECTIVE LASTOBJECTIVE
SETCONSTRAINTS A_CINTAKE03 0 0

SETCONSTRAINTS A_CINTAKE04 -40 0
ADDOBJECTIVE A_CINTAKE04 maximize
RUNFBA CINTAKE04
DELOBJECTIVE LASTOBJECTIVE
SETCONSTRAINTS A_CINTAKE04 0 0

SETCONSTRAINTS A_CINTAKE05 -40 0
ADDOBJECTIVE A_CINTAKE05 maximize
RUNFBA CINTAKE05
DELOBJECTIVE LASTOBJECTIVE
SETCONSTRAINTS A_CINTAKE05 0 0
SETCONSTRAINTS A_CINTAKE06 -40 0
ADDOBJECTIVE A_CINTAKE06 maximize
RUNFBA CINTAKE06
DELOBJECTIVE LASTOBJECTIVE
SETCONSTRAINTS A_CINTAKE06 0 0

SETCONSTRAINTS A_CINTAKE07 -40 0
ADDOBJECTIVE A_CINTAKE07 maximize
RUNFBA CINTAKE07
DELOBJECTIVE LASTOBJECTIVE
SETCONSTRAINTS A_CINTAKE07 0 0
```

## B.6 Artificial reactions

Proportions between metabolites were set by using artificial reactions. For example:

- bio18a: reactants with set of proportions $P_A$ → 1 BIOMASS + byproducts
- bio18b: reactants with set of proportions $P_B$ → 1 BIOMASS + byproducts
- R_A_PHA_SELECTOR (for the wiltype PHA): 1 A_PHA_LFM046 → 1 A_PHA
- R_A_PHA_SELECTOR (for simulation 1): 1.83857 A_PHA_SIM1 → 1 A_PHA
- A_BIOTOT01_u: 7.291 A_PHA + 1 BIOMASS → 1 BIOTOT
- A_CINTAKE02: 0.80 glucose + 0.20 fructose → ∅

All alternative biomass equations were dimensioned for a biomass molar mass of 1000 g mol$^{-1}$ (e.g. the proportions in $P_A$ and $P_B$ and in the byproducts). The correction factor 1.83857 in the reaction R_A_PHA_SELECTOR is to make sure that the artificial metabolite A_PHA has the same molar mass of the wildtype PHA (approx. 165 g mol$^{-1}$), to be able to reuse the A_BIOTOT equations regardless of the PHA composition simulated.

The reaction A_BIOTOT01_u corresponds to a point in the upper surface of the *3*-phenotypic model with a PHA content of approximately 55 %wt. This steady-state was limited in phosphorus and the immediately previous observation of it is set in the reaction A_BIOTOT01_d, a point in the lower surface of the *3*-phenotypic model. A_BIOTOT01_u was the reaction used in simulations where growth and/or PHA content were to be kept constant, by setting the upper and lower bounds of all its alternatives to zero. The reaction A_CINTAKE02 has no products because it is an exchange reaction.