RAFAEL DAVID DE OLIVEIRA

NUMERICAL METHODS FOR METABOLIC NETWORK MODELS

São Paulo

2022

RAFAEL DAVID DE OLIVEIRA

NUMERICAL METHODS FOR METABOLIC NETWORK MODELS

Versão corrigida

Tese apresentada à Escola Politécnica da Universidade de São Paulo para a obtenção do título de Doutor em Ciências.

Área de concentração: Engenharia Química

Orientador: Prof. Dr. Galo Antonio Carrillo Le Roux

São Paulo

2022

Este exemplar foi revisado e corrigido em relação à versão original, sob responsabilidade única do autor e com a anuência de seu orientador.

São Paulo, 31 de Outubro de 2022

Assinatura do autor: Rafael David de Oliveira

Assinatura do orientador:

Catalogação-na-publicação

de Oliveira, Rafael David
Numerical Methods for Metabolic Network Models / R. D. de Oliveira -- versão corr. -- São Paulo, 2022.
125 p.

Tese (Doutorado) - Escola Politécnica da Universidade de São Paulo. Departamento de Engenharia Química.

1.Modelos metabólicos 2.Identificabilidade 3.Controle preditivo baseado em modelo 4.Análise de Balanço de Fluxos 5.13C-Análise de fluxos metabólicos I.Universidade de São Paulo. Escola Politécnica. Departamento de Engenharia Química II.t.

À pessoa mais sábia que conheci em toda minha vida, minha mãe Maria Francisca.

Acknowledgements

Começo agradecendo ao meu orientador Prof. Galo Le Roux por todo o apoio e incentivo durante todos esses anos, e por entender que o trajeto de um doutorando vai muito além da pesquisa que ele realiza. Agradeço também ao Prof. José Gregório por todas as ótimas sugestões e discussões durante essa jornada. I would also like to thank Prof. Mahadevan and Prof. Johannes, for having me in their research groups and for contributing to my education as a researcher.

Agradeço aos meus caros amigos de casa: Leonardo, Gustavo, Rafael (Vaca) e Gabriel (Vila) por todos os debates sobre política, futebol, filosofia e economia. Aos meus caros amigos que fiz nesses seis anos em São Paulo: Priscila (Jesus e Paz), Irena, Zé Otávio, Zé Eduardo, María, Javier, Dielle e Dondon pela amizade e pelo café de todo dia. À Carol pelo apoio em diversos aspectos desse trabalho e pela amizade. Thank to all my colleagues and friends from Toronto and Trondheim that made me feel at home.

Eu gostaria de agradecer à agência de fomento Coordenação de Aperfeiçoamento de pessoal de Nível Superior (CAPES) pelo apoio financeiro nacional e para a realização do doutorado sanduíche.

Agradeço à Tamires pela parceria e por me fazer conseguir superar todas as adversidades durante a elaboração dessa tese. Por fim, gostaria de agradecer à minha família por tornar possível a realização dessa tese e por acreditar em mim sempre.

“The purpose of computing is insight, not numbers.”

(Richard Hamming)

Resumo

A busca por fontes renováveis de energia, materiais e produtos químicos tem sido o foco de muitas políticas internacionais. Bioprocessos são uma promissora alternativa para atingir esses objetivos, entretanto, esses processos devem ser economicamente viáveis para se obter sucesso em um mercado altamente competitivo. A bioprodução precisa ser otimizada tanto em nível celular, como em nível de processo. A biologia de sistemas pode contribuir em ambos os níveis, tornando modelos adequados para a análise de engenharia metabólica, modificando o metabolismo das células, e fornecendo modelos preditivos para otimizar as operações de bioprocessos. Esses métodos computacionais devem ser precisos e confiáveis, portanto, áreas como a Engenharia de Sistemas em Processos (PSE) têm um enorme potencial de contribuição. O objetivo desta tese foi aplicar ferramentas de PSE para aperfeiçoar os métodos existentes em Biologia de Sistemas. Para isso, foram construídas colaborações interdisciplinares para encontrar gargalos nas metodologias existentes. A tese foi dividida em três tópicos/projetos principais. No primeiro, métodos de identificabilidade foram aplicados ao planejamento ótimo de experimentos de carbono marcado para melhorar o processo de estimação de fluxos metabólicos. A metodologia foi aplicada à produção de bioplásticos e biossurfactantes por *Pseudomonas spp.*. O planejamento de experimentos foi utilizado para se determinar o melhor substrato marcado e para reduzir o custo experimental. Foram aplicados três métodos de análise de identificabilidade, baseados na inspeção visual das saídas do modelo, na Matriz de Informação de Fisher e em Componentes Principais. As marcações ótimas para os substratos para cada espécie de *Pseudomonas* foram determinadas e as principais rotas metabólicas foram estimadas. O segundo projeto consistiu em tornar um modelo metabólico dinâmico conhecido como Dynamic flux balance analysis (dFBA) mais adequado para aplicações de controle preditivo baseado em modelos. Os modelos dFBA consistem em um sistema de equações diferenciais ordinárias com um modelo de otimização integrado, portanto, aplicações em controle são desafiadoras. O modelo de otimização foi substituto por um modelo *surrogate* e foi demostrado que esse método pode diminuir consistentemente o tempo computacional e tornar o desempenho do controlador mais confiável. Finalmente, o terceiro projeto também se concentrou em modelos dFBA, mas visando torná-los mais viáveis para aplicações de engenharia metabólica dinâmica. O modelo dFBA foi reformulado como um problema NLP em um único nível. São discutidos os desafios de resolver um problema com restrições complementares e como superá-los. Estudos de caso de controle dinâmico do metabolismo demonstraram o potencial da metodologia desenvolvida. Juntos, esses métodos podem ser aplicados para obter simulações numéricas mais confiáveis e precisas usando modelos de biologia de sistemas.

Palavras chaves: Modelos metabólicos, Identificabilidade, Controle preditivo baseado em modelo, Análise de Balanço de Fluxos, ^{13}C-Análise de fluxos metabólicos

Abstract

The search for renewable sources of energy, materials, and chemicals has been the focus of many international policies. Bioprocess emerged as a powerful alternative to achieve these goals, however, this type of process should be economically feasible to be successful in a highly competitive market. Bioproduction needs to be optimized at the cellular level, as well as at the process level. Systems biology can contribute to that aim, making models suitable for metabolic engineering analysis to modify cells' metabolism, and also providing predictive models to optimize bioprocess operations. These computational methods must be precise and reliable, therefore, fields such as Process Systems Engineering (PSE) have a larger potential to contribute. The aim of this thesis was to apply PSE tools to improve existing systems biology methods. In order to do that, interdisciplinary collaborations were built to find bottlenecks in the existing methodologies. The thesis was divided into three main topics/projects. In the first one, identifiability methods were applied to the optimal design of carbon labeling experiments to improve the metabolic flux distribution estimation process. The methodology was applied to *Pseudomonas spp.* producing bioplastics and biosurfactants under non-growing conditions. Design of experiments was performed to determine the best labeling substrate and also to reduce experimental cost. Three methods for identifiability analysis were applied, based on visual inspection of the response surface of the model output, on the Fisher Information Matrix and a Principal Component-based technique. The optimal labeled substrates for each *Pseudomonas* specie was determined, and the main routes for bioproducts biosynthesis were estimated. The second project consisted in building a dynamic metabolic model known as Dynamic flux balance analysis (dFBA) more suitable for model predictive control applications. dFBA models consist of a system of ordinary differential equations with an optimization model embedded, thus, incorporating them into MPC applications is very challenging. We replaced the embedded optimization model with a simple surrogate model and showed that this can consistently decrease computational time and make the MPC performance more reliable. Finally, the third project also focused on dFBA models but tried to make them more feasible for dynamic metabolic engineering applications. We reformulate the dFBA model as an NLP and showed that the solution of the bi-level optimization problem can be obtained from a single level problem, which is more straightforward to solve. The challenges of solving a complementary constrained problem and the way to overcome that problem are discussed. Case studies of dynamic control of metabolism demonstrated the potential of the developed methodology. Taking all together, these methods can be applied to obtain more reliable and precise numerical simulations using system biology models.

Keywords: Metabolic models, Identifiability, Model Predictive Control, Flux Balance Analysis, ^{13}C-Metabolic Flux Analysis

List of Figures

List of Tables

List of abbreviations and acronyms

6PG	6-phosphogluconate
^{13}C-MFA	^{13}C-Metabolic flux analysis
^{13}C-NMFA	^{13}C-Non-stationary metabolic flux analysis
^{13}C-DMFA	^{13}C-dynamic metabolic flux analysis
AcCoA	Acetyl coenzyme A
ADP	Adenosine diphosphate
ATP	Adenosine triphosphate
Cit	Citrate
CO_2	Carbon dioxide
DA	Direct approach
DAE	Differential-algebraic equation
DHP	Dihydroxyacetone
DO	Dissolved oxygen
DOA	Dynamic optimization approach
DOE	Design of experiments
dFBA	Dynamic flux balance analysis
E4P	Erythrose-4-phosphate
ED	Entner–Doudoroff (pathway)
EMU	Elementary metabolic unity
EMPC	Economic model predictive control
EFM	Elementary flux modes
EFV	Elementary flux vectors

F16P	Fructose 1,6-bisphosphate
F6P	Fructose-6-phosphate
$FADH_2$	Flavin adenine dinucleotide
FBA	Flux balance analysis
pFBA	Parsimonious flux balance analysis
FIM	Fisher information matrix
Fum	Fumarate
FVA	Flux variable analysis
G3P	Glycerinaldehyde-3-phosphate
GAP	Glycerinaldehyde-3-phosphate
G6P	Glucose-6-phosphate
GLX	Glyoxylate
GSM	Genome-scale model
GS-MS	Gas chromatography-mass spectrometry
HAAs	Hydroxyalkanoiloxy-alkanoates
IsoCit	Isocitrate
KKT	Karush-Kuhn-Tucker
LB	Lysogeny broth
LP	Linear programming
LICQ	Linear independence constraint qualification
Mal	Malate
MFA	Metabolic flux analysis
MINLP	Mixed-Integer nonlinear programming

MPC	Model predictive control
MPCC	Mathematical program with Complementary constraints
NADH	Nicotinamide adenine dinucleotide
NADPH	Nicotinamide adenine dinucleotide phosphate
NLP	Nonlinear programming
NMR	Nuclear magnetic resonance
O_2	Oxygen
OAA	Oxaloacetate
ODE	Ordinary differential equations
ODEO	Ordinary differential equation system with embedded optimization
OF	Objective function
PHA	Polyhydroxyalkanoates
P3HB	Poly-3-hydroxybutyrate
PC	Principal components
PCA	Principal component analysis
PCR	Principal component regression
PEP	Phosphoenolpyruvate
PG2	2-phosphoglycerate
PG3	3-phosphoglycerate
PG6	6-phosphogluconate
PHA	Polyhydroxyalkanoate
Pyr	Pyruvate
PLS	Partial least square

PLSR	Partial least square regression
PP	Pentose phosphate (pathway)
PSE	Process systems engineering
QP	Quadratic programming
RHL	Rhamnolipids
Rb5P	Ribose-5-phosphate
Rbl5P	Ribulose-5-phosphate
RMSE	Root mean square error
RMSEP	Root mean square error of prediction
S	Substrate
S7P	Sedoheptulose-7-phosphate
SOA	Static optimization approach
SSE	Sum of squared errors
Suc	Succinate
SucCoA	Succinyl-CoA
X5P	Xylulose-5-phosphate

List of symbols

Chapter 2

v	Vector of metabolic fluxes
$\bar{x}$	Vector of isotopomer fraction
$\bar{P}$	Unimolecular isotopomer transition matrices
$\bar{Q}$	Bimolecular isotopomer transition matrices
x	Vector of cumomer fraction
v^{net}	Net flux
v^{xch}	Exchange flux
v_{free}	Free fluxes
y	Measurements
x_{inp}	Labeling of the substrate
$\sum$	Covariance matrix of the measurements
D	D-criterion
H	Hessian matrix

Chapter 3

v	Vector of metabolic fluxes
lb	Lower bound
ub	Upper bound
x	Concentrations of the external metabolites
u	System manipulated variables
y	Process outputs (measurements)
T_p	Controller prediction horizon

h	Controller sampling time
k_l	Heaviside parameter
F	Feed flow rate
V	Reactor liquid volume
v_g	Glucose uptake flux
v_o	Oxygen uptake flux
K_g	Glucose saturation constant
K_o	Oxygen saturation constant
K_{ig}	Glucose inhibition constant
K_{ie}	Ethanol inhibition constant
v_e	Exchange flux of ethanol
μ	Relative cellular growth
v_z	Uptake flux of xylose
v_e	Exchange flux of ethanol

Chapter 4

v	Vector of metabolic fluxes
x	Concentrations of the external metabolites
h	Time step
lb	Lower bound
ub	Upper bound
λ	Multipliers of equality constraint
α	Multipliers of inequality constraint
ρ	Penalization parameter

Φ	Objective function
v_g	Glucose uptake flux
v_o	Oxygen uptake flux
K_g	Glucose saturation constant
K_o	Oxygen saturation constant
K_{ig}	Glucose inhibition constant
K_{ie}	Ethanol inhibition constant
v_e	Exchange flux of ethanol
μ	Relative cellular growth
v_z	Uptake flux of xylose
v_e	Exchange flux of ethanol
ϵ	Flux constraint parameter
γ	Element size change constraint
t_{reg}	Time of regulation
F	Feed flow rate
O_{sat}	Oxygen saturation concentration

Contents

1 Introduction

The search for a sustainable economy has driven the scientific community to look for replacements for fossil-derived products. Bioproducts are, in this context, the natural replacement candidates. Special attention has been dedicated to producing fuels, material, and chemicals by microorganisms (ZETTERHOLM et al., 2020). That fact results from the high flexibility of compounds that can be produced by cells (LEE; KIM, 2015). As examples of commercialized bioproducts, it is possible to mention ethanol, isobutanol, acetone, lactic acid, and polyhydroxyalkanoates (PHA) (GUSTAVSSON; LEE, 2016).

Although it is possible to mention some successful cases of industrialized biocompounds, for most fossil-derived products, the equivalent bioproduct is far more expensive (ZETTERHOLM et al., 2020). The cost of the process is usually an impediment, and in some cases, a new biological route must be developed. Biological systems are complex, and the complete knowledge of the cell's functionality is not yet uncovered. However, in the past decades, a variety of new tools have emerged with the development of fields like systems biology (KITANO, 2002) and metabolic engineering (STEPHANOPOULOS, 1994). These fields have in common the use of mathematical models to analyze the cell with a holistic view.

The concept of microbial cell factories can be explained from an analogy between the cells and a chemical plant. As a chemical plant can be modified in order to produce different products, in the same way, cells can be modified to generate a new bioproduct. Molecular biology techniques for genetic modifications are now routine in research labs around the world. The process of selecting a host organism, designing pathways, and genetic engineering this host to produce a particular bioproduct, it is known as strain optimization (GUSTAVSSON; LEE, 2016). The field of metabolic engineering emerged as a rational way of performing strain optimization by the utilization of mathematical tools (STEPHANOPOULOS, 1994). The goals of metabolic engineering usually are: improving substrate utilization; enabling a wide range of carbon source utilization; increasing product tolerance; removal of feedback inhibition; and robustness.

Systems Biology is an interdisciplinary field that seeks to represent complex biological systems by mathematical models (PALSSON, 2015). By applying a holistic view, systems biology aims to find emergent properties of cells, tissues, and organisms. In particular, the application of systems biology models to the study of cellular metabolism has significantly increased the understanding of the cellular functions (NIELSEN, 2017).

Nielsen and Keasling (2016) and Lee and Kim (2015) detailed some metabolic engineering and systems biology achievements and the role played by mathematical models in each of them. In all cases, metabolic models had an important role in the development of the new strains by giving genetic modification targets, metabolic flux maps, or by optimizing the bioprocess operation. In the next section, a general introduction will be presented to a class of models (Metabolic Network models) widely applied in Systems Biology and Metabolic Engineering.

Metabolic Network Models

There are diverse processes going on at the same time inside a cell, trying to understand all these processes is a hard task to be achieved. A popular way to look at and to organize these processes is to separate them into networks. A network consists of nodes that represent compounds and links that are chemical transformations that relate these compounds. In this sense, for each specific microorganism, it is possible to formulate a signaling network, regulatory network, metabolic network, etc. The most studied network of a cell is the metabolic network and consequently it is available for a variety of species (FANG et al., 2020). Metabolic network consists of metabolite nodes that are connected by enzymatic reactions. A recent and more advanced class of metabolic networks is the genome-scale metabolic networks, where the enzymatic reactions are related with the respective genes. These networks have become a powerful tool to give a systemic view of metabolism.

The process of building a metabolic network is called reconstruction. The reconstruction of a metabolic network consists in compiling a large amount of data like proteomics, metabolomics and fluxomics and try to put all of this together in a matrix called stoichiometric matrix (PALSSON, 2015). Each row of a stoichiometric matrix represents a metabolite and each column an enzymatic reaction. A special reaction in this matrix is a pseudo biomass reaction, where some pseudo reactions of basic constituents of the cell's biomass are lumped together. Once the stoichiometric matrix is built, it is possible to develop models based on it.

Different classes of metabolic models can be formulated, here they are classified as steady-state, carbon flux, and dynamic models. The steady-state models have as the main assumption that the metabolic fluxes and metabolite pool sizes do not change with time. Usually, inside a cell, there are many more reactions than metabolites which implies that the system of equations formed by the stoichiometric matrix has multiple solutions, in other

words, multiple possible flux distributions. For the few cases of small networks when this is not true, the technique Metabolic Flux Analysis (MFA) can be applied, where the external fluxes are measured and the internal fluxes are estimated using the stoichiometric matrix (STEPHANOPOULOS, 1994). MFA has historical importance because it was the first metabolic model to be used (STEPHANOPOULOS, 1994), however, its applications are limited.

For the other cases with multiple possible solutions, two kinds of models can be applied: biased or unbiased. Elementary Flux Modes (EFM) is an unbiased method because it does not require any further assumption about the cellular metabolism behavior. EFM finds an algebraic basis for the flux space, in other words, EFM finds pathways that keep the cell in steady-state conditions (SCHUSTER; HLGETAG, 1994). EFM is very useful to compare product yields for different metabolic pathways and find correlations between reactions. However, the application of EFM to predict a flux distribution is not straightforward, because it does not give a single solution. Flux Balance Analysis (FBA) is a biased method, that assumes an objective function for the cell in order to find a flux distribution. Typically it is assumed that the cells have the objective of maximizing biomass yield (ORTH et al., 2010). The FBA has the advantage of giving a single solution of the flux distribution and being much easier to interpret, however equally likely solutions are common and the assumption of maximizing biomass yield is not always valid (ORTH et al., 2010). FBA is very useful to predict promising gene deletions and how the insertion of new pathways influences on metabolism.

The main limitation of steady-state models is the availability of data. The carbon flux models try to overcome this difficulty by using more information about the interior of the cell. The core of this method is to use a labeled substrate (usually carrying carbon isotope 13) and supply this substrate to the cells' medium. As the cell consumes it, the labeling is spread out in all metabolites until it finally comes out in the bioproducts. The labeling pattern is usually measured in amino acids and bioproducts by techniques such as Gas Chromatography-Mass Spectrometry (GS-MS) and Nuclear Magnetic Resonance spectroscopy (NMR) (ANTONIEWICZ, 2015). The fate of all atoms in the metabolic network must be known to formulate a model that simulates the labeling spread in cells. As the measured patterns are a function of the metabolic fluxes, a flux distribution can be estimated using carbon flux models and the labeling measured data. When this technique is applied in metabolic and isotopic steady-state it is called ^{13}C- Metabolic Flux Analysis (^{13}C-MFA) (WIECHERT, 2001). Isotopic steady state takes place when the labeling does not change with time. A much more complex technique is applied in the isotopic non-stationary condition,

which is called ^{13}C- Non-stationary Metabolic Flux Analysis ^{13}C-NMFA. ^{13}C-NMFA has required advanced analytical techniques to perform fast analysis and is computationally costly, however, it enables the estimation of a higher amount of metabolic fluxes with more precision (NöH et al., 2006). Finally, the technique that is performed outside the metabolic and isotopic steady state is called ^{13}C dynamic metabolic flux analysis (^{13}C-DMFA), however, only exploratory work has been reported because of its complexity (ANTONIEWICZ, 2015).

Although steady-state models had achieved success in many applications, they have serious limitations, such as the lack of representation of metabolite concentrations and enzymatic regulation (SRINIVASAN et al., 2015). These features are key concepts to analyze real industrial bioprocess cases, which are essentially dynamic. Kinetic models can represent the biological dynamic states with more physical precision. Nevertheless, building kinetic models is not yet common because of many obstacles such as (SAA; NIELSEN, 2017): lack of available data; parameter identifiability and estimability; and model non-linearity. Finally, there are also hybrid models, that use stoichiometric models to represent the internal metabolism and kinetic expression of key uptake rates. Examples of hybrid models are Dynamic Flux Balance Analysis (dFBA) (MAHADEVAN et al., 2002) and Cybernetic models (RAMKRISHNA; SONG, 2018).

Numerical challenges in metabolic network models

Behind the application of metabolic network models to bioprocess development there are numerical methods used to solve these models. Usually, these methods are implemented with the aid of a software where the user only has access to the interface. The numerical methods applied to solve these models are far from being free of errors and pitfalls, that are usually reported in the literature.

Chindelevitch et al. (2014) described some numerical issues in the toolbox for FBA simulation and proposed an exact arithmetic toolbox to deal with these issues, and a discussion followed from this work (EBRAHIM et al., 2015; CHINDELEVITCH et al., 2015). In ^{13}C-MFA, Follstad and Stephanopoulos (1997) pointed out that some simplification on reversible reaction could lead to flux misestimations and Winden et al. (2001b) described pitfalls associated to commonly applied model reduction techniques. Kohlstedt and Wittmann (2019) showed that some assumptions in the design of experiments could lead to flux identifiability problems. Vasilakou et al. (2016) defined challenges for dynamic metabolic modeling. Some of them

are computational/numerical issues like developing rigorous dynamical systems theory and new computational tools for parameter exploration and identification. New methods must be developed and, as already pointed out, Process Systems Engineering (PSE) has a tremendous potential to contribute (MARANAS et al., 2003).

1.1 Aim and outline of the thesis

This thesis aimed to study numerical methods applied to metabolic network models. This "learning by doing" process revealed some drawbacks to existing methodologies. Therefore, the aim was to develop new methodologies and improve existing ones.

The thesis is separated into three sections/chapters, each one with a different application of metabolic models. Chapter 2 describes the application of identifiability analysis to the design of carbon labeling experiments. The case study is the simultaneous production of PHA and rhamnolipids by *Pseudomonas aeruginosa*. ^{13}C-MFA in non-growing conditions can be challenging, and identifiability analysis had to be applied to overcome the scarcity of available data. Chapter 3 presents a surrogate model approximation of FBA models. The surrogate FBA can replace the original FBA optimization problem and avoid bi-level optimization problems that arise in some dFBA applications and that be computationally costly. Two case studies using the metabolic network of *Saccharomyces cerevisiae* were solved: A model Predictive Control (MPC) of a bioreactor, and a parameter estimation problem. Finally, Chapter 4 presents an NLP formulation for the dFBA models that replaces the FBA optimization by the first-order optimality conditions. Some challenges in solving this problem are highlighted and some solutions are proposed. Five case studies are presented and the advantages and drawbacks of the method are discussed.

2 Identifiability of metabolic flux ratios on carbon labeling experiments

Metabolic Engineering is a field that aims to assist in the development of new strains for the sustainable production of valuable bioproducts. In essence, it consists in applying mathematical models to guide genetic modification in cells in order to obtain higher yields and productivity (SAUER, 2006). Among the tools that have been developed in the field, ^{13}C-Metabolic Flux Ratio Analysis (^{13}C-MFA) emerged as a powerful tool to describe the metabolism of a cell by estimating the flux distribution of metabolites (SAUER, 2006). Figure 1 illustrates how this method works; first a 13C labeled substrate is used in the culture medium (e.g. [6-^{13}C-glucose]), then the labeling goes through the internal metabolites until it reaches a bioproduct. The labeling pattern in the bioproduct is a function of the metabolic pathway that was used by the cell, and each pathway shuffles the labeling differently. Finally, the labeling in the bioproduct is measured by techniques such as Gas Chromatography-Mass Spectrometry (CG-MS) and Nuclear magnetic resonance (NMR). Using a metabolic network model, an estimation problem is formulated and solved and the flux distribution inside the cell estimated. Determining precisely which fluxes can be estimated from a given experiment architecture (i.e. the particular substrate labeling and bioproducts measurements) is crucial due to the high experimental cost associated with carbon labeling experiments.

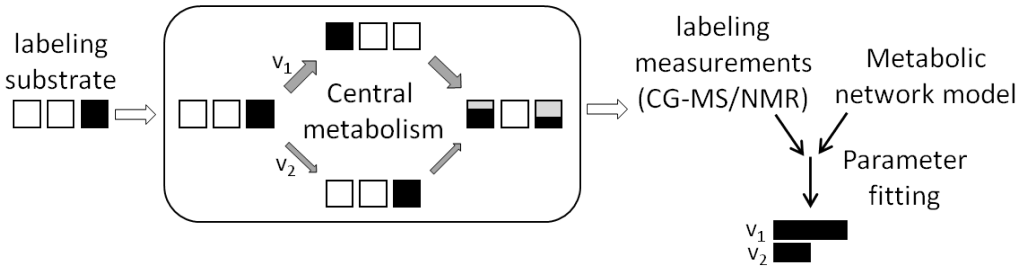


Figure 1 – ^{13}C-MFA method.

In this chapter, the mathematical formalism of ^{13}C-MFA is presented with a focus on the importance of the design of labeling experiments and identifiability analysis to reduce experimental costs and the estimation uncertainty. Three different methods for identifiability analysis were applied to study the central metabolism of *Pseudomonas* during the production of polyhydroxyalkanoates (PHA) and/or rhamnolipids (RHL). The production of these bioproducts takes place on non-growth conditions, where only a few labeling

measurements are available to perform flux estimation, making it a challenging problem to solve.

Remark. *The results presented here are based on the works Oliveira et al. (2021b) and Oliveira et al. (2021c). Three identifiability analyses were applied in different steps of the development of this doctoral thesis with different aims. Here, all of these analyses were put together in an effort to represent all the knowledge developed by the research group for applying ^{13}C-MFA on Pseudomonas.*

2.1 Research background

^{13}C-Metabolic Flux Analysis

Isotopomers

In order to describe the pattern of labeling of each metabolite in a cell during carbon labeling experiments, the concept of isotopomer is applied. Isotopomers are isomers with isotopic atoms, having the same number of each isotope of each element but differing in their positions. An example is given in Figure 2 for a molecule of three atoms. The difference of the isotopomers is due to where the labeled atoms are placed (filled circles). The position isotopomers is the set of all possible isotopomers. A subset of isotopomers is the mass isotopomers where the difference is made by the isotopomer's weight.

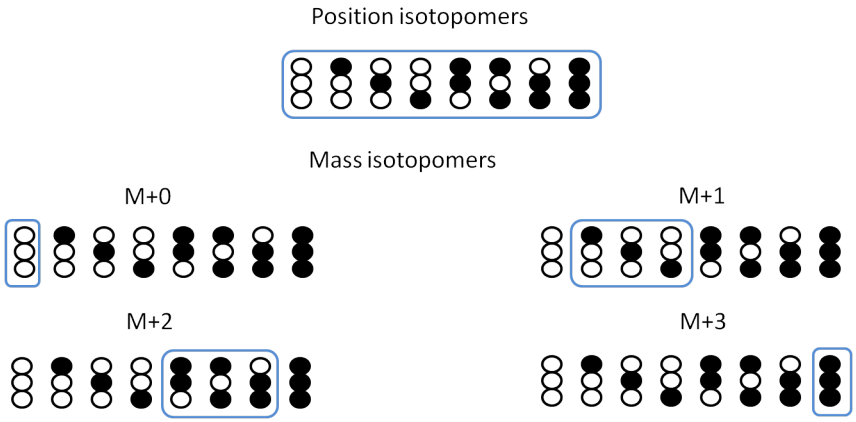


Figure 2 – Comparison between position and mass isotopomers. Filled circles represents isotopic atoms.

Modeling and simulation of carbon labeling experiments

The mathematical model is a crucial part of ^{13}C-MFA method. The model needs to describe the carbon labeling patterns in each metabolite as a function of the metabolic flux distribution inside the cell. In order to do that the metabolic network of the metabolism and the atom transition map of the enzymatic reactions must be known *a priori*.

Zupke and Stephanopoulos (1994) were the first to propose a mathematical model to account for labeling spreading in metabolic networks. The model has the form:

$$\bar{f}(v,\bar{x}) = \frac{1}{2}\bar{x}^T \cdot (\sum_i \vec{v} \cdot \vec{\bar{Q}}_i + \overleftarrow{v} \cdot \overleftarrow{\bar{Q}}_i) \cdot \bar{x} + (\sum_i \vec{v}_i \cdot \vec{\bar{P}}_i + \sum_i \overleftarrow{v}_i \cdot \overleftarrow{\bar{P}}_i) \cdot \bar{x} = 0 \tag{2.1}$$

where v is a vector of metabolic fluxes; $\bar{x}$ is a vector of isotopomer fraction; $\bar{P}_i$ are unimolecular isotopomer transition matrices; $\bar{Q}_i$ are bimolecular isotopomer transition matrices. The Equation 2.1 is linear with respect to the fluxes (v), but it is bilinear with respect to isotopomer fractions ($\bar{x}$). These bilinear terms come from the bimolecular reactions in metabolic networks, and they introduce difficulties to solve the model. Solving systems of nonlinear equations using a numerical method can be time-consuming and a source of uncertainties in the solution (BEERS, 2007). Since the work of Zupke and Stephanopoulos (1994) until now, a series of transformations of variables were proposed in the literature to obtain an explicit solution for the isotopomer balance equations:

- Cumomers: sums of the original isotopomer variables (WURZEL; GRAAF, 1999);
- Bondomers: the carbon bonds are used to make the transformation (WINDEN et al., 2002);
- Elementary Metabolic Units (EMU): eliminates isotopomer fractions that are not relevant to the problem. (ANTONIEWICZ et al., 2007);
- Fluxomers: combine the flux and labeling fraction variables (SROUR et al., 2011);

The cumomer variable transformation was the first to be proposed and together with the EMU, are the most applied transformations (ANTONIEWICZ, 2015). Recently, the cumomer transformation was applied in a ^{13}C-MFA toolbox developed in our research group (OLIVEIRA, 2018). When the cumomer transformation is applied, Equation 2.1 becomes:

$$
\begin{aligned}
1 &= {}^{0}x \\
0 &= {}^{1}A(v) \cdot {}^{1}x + {}^{1}b(v) \\
0 &= {}^{2}A(v) \cdot {}^{2}x + {}^{2}b(v, {}^{1}x) \\
0 &= {}^{3}A(v) \cdot {}^{3}x + {}^{3}b(v, {}^{1}x, {}^{2}x) \\
&\vdots
\end{aligned}
\tag{2.2}
$$

where ${}^{n}x$ are vectors of cumomer fraction of weight n; A and b are cumomer transition matrices and vectors, respectively. The system of Equations 2.2 is linear if solved in sequential order of the cumomer weights (n). Therefore, it is much faster and more precise to solve. Equation 2.2 can be used to calculate a cumomer fraction x from a given flux distribution v.

Bidirectional reactions are crucial in ^{13}C-MFA (WIECHERT; GRAAF, 1997; WIECHERT et al., 1997; WURZEL; GRAAF, 1999; MöLLNEY et al., 1999). Normally, one flux variable is associated with one reaction direction, however, this would imply in non-identifiability of both revere and forward fluxes and, as consequence, cause problems of convergence in the estimation process (WIECHERT; GRAAF, 1997). To avoid some of these problems Wiechert and Graaf (1997) suggested a parameter transformation of bidirectional fluxes as:

$$
v^{net} = \vec{v} - \overleftarrow{v} \tag{2.3}
$$

$$
v^{xch} = min(\vec{v}, \overleftarrow{v}) \tag{2.4}
$$

where v^{net} represents the net flux in the reaction, and can be positive or negative depending on the reaction direction; v^{xch} represents the exchange in the reaction being a positive number. For unidirectional reactions, $v^{xch} = 0$, and for fast equilibrium reactions, $v^{xch} \Rightarrow \infty$. In order to reduce the dimension of fluxes to be estimated from labeling data, the metabolic fluxes (v) are parametrized using the metabolite balances ($S\,v = 0$) and calculated in terms of the so-called free fluxes ($v = g(v_{free})$). Finally, it is possible to compute the sensitivities of the labeling of the measured components with respect to the free fluxes in an explicit form differentiating the system of equations 2.2:

$$
f(x, x_{inp}, v_{free}) = 0, \quad y = h(x) \tag{2.5}
$$

$$
\frac{\partial y(x, v_{free})}{\partial v_{free}} = \frac{\partial y(x)}{\partial x} \cdot \frac{\partial x(v_{free})}{v_{free}} \tag{2.6}
$$

where f is System of Equations 2.2, x_{inp} is the labeling of the substrate and y the labeling of the measurements.

Optimal Design of labeling experiments and identifiability analysis

The utilization of a mathematical model for designing the labeling experiment is called Design Optimal of Labeling Experiment that we are going to call just Design of Experiments and use the acronym DOE (ANTONIEWICZ, 2013). DOE usually applies mathematical model simulations to choose the pattern of substrate labeling and the labeling measurements that would minimize the experimental costs while allowing for the estimation of the aimed set of fluxes with minimal uncertainty. The number and type of measurements can directly impact the cost of the labeling experiment. Furthermore, the choice of the substrate labeling pattern can also have an impact on experimental cost, as an example, 1 g of [1-^{13}C]Glucose costs around R$1625 (SIGMA-ALDRICH, 2021).

A field related to DOE is identifiability analysis, that consists of a group of methods that are used to determine how well the parameters of a model can be estimated by a given quantity and quality of experimental data (MCLEAN; MCAULEY, 2012). In the context of metabolic flux estimation, ^{13}C-MFA does not always provide a unique flux distribution for a given set of labeling measurements. When a metabolic flux cannot be uniquely estimated from the available experimental data, that flux is said to be non-identifiable. Non-identifiability can be classified as structural or practical. Structural non-identifiability takes place if for multiple values of the flux parameters the same labeling pattern would be obtained. Examples of methods for structural identifiability analysis are Taylor series expansion, similarity transformations and generating series (MCLEAN; MCAULEY, 2012). Structural analysis just analyses the structure of the model and does not take into account the experimental errors and the experimental setting. On the other hand, practical identifiability analysis considers the available measurements and the noise in the system. Practical identifiability methods are based on Fisher Information Matrix (FIM), visual inspection of the response surface of the model output, Monte Carlo simulations and Correlation matrix (MCLEAN; MCAULEY, 2012).

The process system engineering community has given important contributions to the study of the identifiability of nonlinear model parameters (BARD, 1974; MCLEAN; MCAULEY, 2012). Wiechert (1995) was one of the first to study the subject. He performed a structural analysis using a Gröbner Basis algorithm. Winden et al. (2001a) applied a structural identifi-

ability analysis using a symbolical algorithm to reduce the metabolic network to a minimal set of reactions that can possibly be estimated. Isermann and Wiechert (2003) developed a general theory of structural flux identifiability of carbon labeling experiments. Chang et al. (2008) developed an integer linear programming to compute the optimal measurements on carbon labeling experiments. Kappelmann et al. (2016) performed a structural identifiability analysis of anaplerotic reactions in *Corynebacterium glutamicum* by evaluating the remaining degrees of freedom of the system by computing the rank of the Jacobian matrix. Möllney et al. (1999) applied a practical identifiability analysis to the DOE. They computed the Fisher Information Matrix and used the D-criterion to determine the best tracers. Finally, Theorell et al. (2017) developed a practical identifiability approach that compute Bayesian confidence intervals using Markov Chain Monte Carlo technique.

Taking all together, most of the work in literature focuses either on structural approaches or in complex algorithms to study the identifiability of metabolic fluxes on carbon labeling experiments. Here, we will focus on practical identifiability analysis of a DOE problem. The cost of experiment will be briefly addressed when different mixtures of labeled substrates will be considered. The type and number of measurements were not taken into account in this project because the measurements were restricted to those data available in our research group laboratory.

Polyhydroxyalkanoates and Rhamnolipids metabolism in *Pseudomonas aeruginosa*

Among the candidate bioproducts presenting high potential to replace petrochemical compounds, two have received special attention: Polyhydroxyalkanoates (PHA) and Rhamnolipids (RHL) (RANDHAWA; RAHMAN, 2014; RAZA et al., 2018). PHA are intracellular bacterial polyester granules accumulated more expressively when cell growth is limited by an essential nutrient such as nitrogen or phosphorus, and carbon is in excess. Those polyesters act as a carbon and energy reserve and can represent up to 80% of the cell final dry weight (RAZA et al., 2018). PHA are thermoplastics materials, therefore they are sustainable substitutes to petroleum-based polymers (RAZA et al., 2018). Rhamnolipids are glycolipid biosurfactants that are secreted in the medium by cells and, similarly to PHA, they are mainly synthesized under nutrient limitation conditions. The applications of rhamnolipids are vast and present in different fields like pharmaceutical, cosmetics, food, laundry,

agriculture and bioremediation (RANDHAWA; RAHMAN, 2014). In spite of the wide range of applications and positive environmental impact related to these bioproducts, their production cost is still higher than petroleum derivatives. PHA and Rhamnolipids can be produced simultaneously in Pseudomonas aeruginosa (HORI et al., 2002). A broader understanding of the metabolism in the biosynthesis of these bioproducts is necessary to explore approaches to achieve the maximum yields. PHA and RHL have a common precursor in *P. aeruginosa* metabolism, (R)-3-hydroxyalkanoic acids, channeled from the *de novo* fatty acids biosynthesis pathway when glucose is used as carbon source (REHM et al., 2001). This is a cyclic pathway that extends fatty acids with two carbons derived from acetyl-CoA after each round. In a reaction catalyzed by PhaG, the intermediate (R)-3-hydroxyalkanoil-ACP is converted into (R)-3-hydroxyalkanoil-CoA, the substrate for PHA polymerization by PHA synthase. (R)-3-hydroxyalkanoil-ACP is also the substrate for the reaction catalyzed by RhlA leading to the synthesis of hydroxyalkanoiloxy-alkanoates (HAAs) the precursor in rhamnolipids biosynthesis (ZHU; ROCK, 2008). The glucose catabolism by Entner-Doudoroff (ED) pathway supplies acetyl-CoA and redox power (NADPH) for PHA and RHL biosynthesis. ED can operate in a cyclic or non-cyclic mode in *Pseudomonas* (NIKEL et al., 2015). Kohlstedt and Wittmann (2019) estimated null flux through the PP pathway (oxidative branch) and a cyclic operation of ED pathway in the human pathogen *P. aeruginosa* PAO1 during the exponential growth phase. *P. aeruginosa* LFM634 (like *P. aeruginosa* PAO1) does not present the oxidative branch of PP, due to the absence of gene *gnd*, making ED the main pathway for NADPH supply.

Carbon labeling experiments in *Pseudomonas aeruginosa*

Proteinogenic amino acids are the target of the labeling measurements in the majority of ^{13}C-MFA reports (MCKINLAY et al., 2014). However, since proteinogenic amino acids are poorly synthesized at the stationary growth phase, the elucidation of the metabolism under non-growing conditions is a clear demand (MCKINLAY et al., 2014). Choi et al. (2011) investigated the metabolic pathway for the biosynthesis of RHL and PHA by *P. aeruginosa* from [1-^{13}C] octanoic acid using ^{13}C–NMR analysis. Riascos et al. (2013) estimated the flux ratio into PP and ED node in *Pseudomonas sp.* producing PHA based on experiments with mixtures of [U-^{13}C] glucose (fully labeled glucose) and [U-^{12}C] glucose (natural glucose).

To the readers interested in ^{13}C-MFA studies of others *Pseudomonas* species the review published by Mendonca et al. (2020) is recommended.

This work aimed to understand the activity of central metabolic pathways in *P. aeruginosa* LFM634 when producing PHA and RHL. Identifiability analysis and a design of experiments were carried out, followed by the labeling experiments.

2.2 Methodology

Identifiability of metabolic fluxes by ^{13}C-MFA

Visual inspection of the response surface of the model output

Visual inspection of the response surface of the labeling simulations for different values of metabolic fluxes can provide useful information on the fluxes identifiability for a given experimental setup. This method consists in performing a series of simulations for different values of metabolic fluxes and observing the labeling of metabolites available as measurement for a given experimental setup. The method is simple to implement, easy to analyze, and does not require an initial guess of the flux distribution. One drawback of the method is the high computational effort to perform the simulations needed to cover all the flux domain. Furthermore, the method is restricted to small models because of the impossibility of visualizing a high-dimensional flux space.

For the analysis of the central metabolic network of *P. aeruginosa* LFM634 we solved the model for every value of the metabolic fluxes in a equidistant grid of 10 by 10 grid. The labeling of the metabolites that can be measured was plot as a response surface.

Fisher Information Matrix based analysis

One criterion used in DOE is to choose the experiments which will provide the smallest variance of the estimated parameters. A classical way to access this information is to compute an estimate of the Fisher Information Matrix (FIM) that gives the variance of the expected parameters values. The FIM can be calculated by inverting the covariance matrix of the parameters, that in turn can be approximated by the sensitivities (Equation 2.6) as follows:

$$FIM(v^0, u) = Cov(v^0)^{-1} \approx \frac{\partial y(x)}{\partial v_{free}}^T \cdot \Sigma^{-1} \cdot \frac{\partial y(x)}{\partial v_{free}} \tag{2.7}$$

where v_0 is the vector of fluxes assumed in the analysis; u the selected labeled substrate mixture; y are the measured mass isotopomers; $\sum$ is the covariance matrix of the measurements. The FIM is a square matrix with the size of the number of degrees of freedom (free fluxes). For this reason different criteria were formulated in order to synthesize the FIM information (MCLEAN; MCAULEY, 2012). The D-criterion can be calculated by the determinant of the FIM as follows:

$$D = det(FIM) \tag{2.8}$$

Maximizing the D-criterion is equivalent to minimizing the volume of the joint confidence region of the parameters (BARD, 1974).

Principal component based analysis

The well-known principal component analysis (PCA) is applied to identify linear relations between metabolic fluxes in ^{13}C-MFA. Also, a methodology recently developed to obtain sparse PCA components is applied (NAKAMA et al., 2020).

Principal component (PC) was applied to assess the identifiability of the parameters of the nonlinear model and select a subset of parameters that can be estimated with high accuracy (VAJDA et al., 1989). This method consists in computing the Hessian matrix (H) obtained by the Gauss-Newton approximation, using normalized sensitivities. Then, H is decomposed into PC, by the Singular Value Decomposition (SVD) as follows:

$$\frac{\partial y(x, v_{free})}{\partial v_{free}} = U\Sigma V^T \tag{2.9}$$

$$H = \frac{\partial y}{\partial v_{free}}^T \cdot \frac{\partial y}{\partial v_{free}} = (U\Sigma V^T)^T \cdot (U\Sigma V^T) = V\Sigma^2 V^T \tag{2.10}$$

where V is the matrix of PCs (eigenvectors of H) and Σ^2 the matrix with the eigenvalues of H. The PCs of H represent linear combinations of the original parameters, which may indicate dependencies between parameters . However, because these matrices are normally dense, finding combinations with a reduced number of relevant parameters can be challenging. Recently, Nakama et al. (2020) developed a method to compute sparse PC that are orthogonal

to the components associated with small eigenvalues, which can be specifically associated to the large variance of the parameters. In this work, both PC and sparse PC were computed to evaluate the identifiability of metabolic fluxes.

Carbon labeling experiment

For this work, a partnership with Prof. Dr. José Gregório (ICB-USP) and the MSc. student Vânia Novello (ICB-USP) who performed all the experiments with labeled substrate was established. The microorganism used in the experiments was *Pseudomonas aeruginosa* LFM634 (PEIXOTO, 2008). This strain was cultivated in two different media. Mineral medium was used to produce RHL and PHA. The LB medium (Lysogeny broth) is a rich medium suitable for the growth and maintenance of various microorganisms. The LB medium was used for inoculum preparation to obtain a high concentration of cells to be further transferred to the mineral medium (RAMSAY et al., 1990). The carbon source of the two experiments was a mixture of ^{13}C glucose and natural glucose, 20% (w/w) of [U-^{13}C]Glucose in the first experiment and 55% (w/w) of [6-^{13}C]Glucose in the second experiment. Nitrogen source was $NaNO_3$ and the carbon/nitrogen mass ratio (w/w) used was C/N=45. A previously prepared culture was inoculated to the mineral medium in a ratio of 10% (v/v). Batch cultivations for 48h and 24h in 250 mL Erlenmeyer flasks containing 50 mL medium were performed in a rotary shaker (30 ℃ and 150 rpm).

The analysis of the 3HA and the 3HAA were done by gas chromatography-mass spectrometry (GC-MS) after their derivatization by propanolysis (RIIS; MAI, 1988), generating 3HA methyl esters. There are three typical fragments in 3HA propyl esters (Appendix: Figure 33). The fragments m/z=[M-59] are specific for each 3HA monomers and include carbons derived from all Acetyl-CoA molecules used in the 3HA biosynthesis and none from the propanol used in the derivatization reaction. The fragment m/z=131 is generated by α cleavage at the hydroxyl functional group (LEE; CHOI, 1995) and includes three carbons derived from biosynthesis and three carbons from the derivatization agent (propanol). The fragment m/z=89 is derived from a McLafferty rearrangement of the three first carbons in the 3HAs (LEE; CHOI, 1995). Therefore, fragments m/z=89 and m/z=131 are common for all 3HA and contain carbons derived from one and a half Acetyl-CoA molecules.

^{13}C-Metabolic flux ratio analysis

A program developed in MATLAB (Version R2013a, MathWorks) was used to perform the carbon labeling experiment simulations, DOE and metabolic flux estimation by ^{13}C-MFA (OLIVEIRA, 2018). The problem of estimating metabolic fluxes ratios was formulated as a nonlinear constrained least squares problem, as follows:

$$min_v \sum_j \left(\frac{x_j^c(v_{free}) - x_j^m(v_{free}}{\sigma_j}\right)^2 \quad \text{subject to:} \quad \text{Equation(2.2)} \quad \text{and} \quad v \geq 0 \qquad (2.11)$$

where v_{free} is the vector of the free fluxes, and x is the cumomer fractional labeling. Indexes c and m indicate calculated and measured quantities respectively; σ corresponds to the standard deviation of the measurements. The objective function is non-linear, thus possibly giving rise to non-convexity. Therefore, a numerical method is necessary to solve this estimation problem; the function *fmincon* was used to find the objective function minima, by applying a Trust-Region method. The parameter optimization was initialized from different starting values and the solution corresponding to the lowest objective function value was selected.

The algorithm for evaluating analytical sensitivities of the labeling state with respect to the fluxes was implemented in the software. For all flux data, 95% confidence intervals were calculated as in Wiechert et al. (1997) applying Fisherian statistics. The simulated mass isotopomers were corrected taking into account the natural abundance of all stable isotopes (WINDEN et al., 2002).

2.3 Results

Case Study: Glucose metabolism of *Pseudomonas aerugionosa* under non-growth conditions

As mentioned at the end of the introduction section of Chapter 2, during the development of this thesis, multiple analysis were performed to evaluate the estimation of metabolic fluxes on *Pseudomonas aerugionosa*. The aim of these analysis was to determine the identifiability of the metabolic fluxes of *Pseudomonas aerugionosa* under non-growth conditions consuming glucose and producing PHA and/or RHL. Usually, the ^{13}C-MFA method is applied under growth conditions and the labeling measurements of amino acids are available (SAUER,

2006). However, PHA and RHL production by *P. aeruginosa* mainly occurs under non-growth conditions and, in this case, only the measurements of 3HAs/3HAAs mass isotopomer are available (RIASCOS et al., 2013; OLIVEIRA et al., 2021b; OLIVEIRA et al., 2021c).

The metabolic network used here (Figure 3) contains the Pentose Phosphate (PP) pathway; the Entner-Doudoroff pathway (ED) that can operate as a linear or cyclic mode; The cyclic mode of the ED pathway is highlighted in Figure 3, where glucose 6-phosphate (G6P) is oxidized to 6-phosphogluconate (6PG), and it is converted into glyceraldehyde 3-phosphate (G3P) and pyruvate (Pyr). G3P is recycled to fructose 6-phosphate which is then converted back to to G6P. The network also comprises a pathway to 3HAAs and 3HAs biosynthesis, where two molecules of acetyl-CoA are condensed into one molecule of 3HAAs or 3HA. These mass isotopomers can be measured in GC-MS analysis and these are the only measurements available from the labeling experiments. Metabolites that are typically converted by fast exchange reactions are lumped in a unique pool (WIECHERT, 2007). Linear reactions are lumped, and fluxes related to growth conditions were set to zero, which includes the TCA cycle reactions. The complete set of reactions and atom transitions is presented in the Appendix (Table 13) .

The network (Figure 3) has 15 reactions (including three reversible reactions from the PP pathway), 9 internal metabolites, and glucose uptake is fixed at the unity. Consequently, the system has five degrees of freedom, that is, five free fluxes. The chosen free fluxes are: the net flux of the oxidative PP pathway (PP_{oxi}); the net flux of the cyclic mode of the ED pathway (ED_{cyc}); and the exchange fluxes of the PP pathway (PP_{xch_1}, PP_{xch_2}, and PP_{xch_3}).

Different strains of *P. aeruginosa* can have different metabolic networks. A common variation on the metabolism of *P. aeruginosa* is the lack of the oxidative branch of the PP pathway, due to the absence of gene *gnd* encoding for 6-phosphoglunate dehydrogenase. The visual inspection of the response surface of the model output and the FIM based identifiability analysis were performed based on the draft genome sequence of *Pseudomonas aeruginosa* LFM634 where the the *gnd* gene was not identified. All the experiments presented in this thesis were performed using this strain. However, because strains of *Pseudomonas spp* that contains the gene *gnd* was also identified in our lab, the PCA based identifiability analysis also considered the case where the oxidative branch of PP is active (PP_{oxi}).

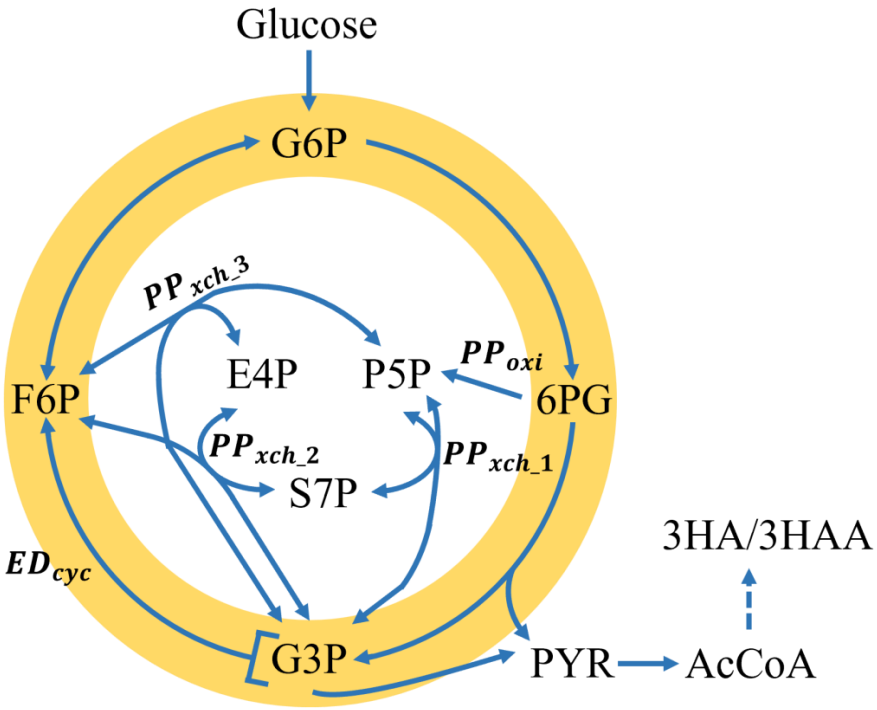


Figure 3 – *Pseudomonas spp.* central glucose metabolism metabolic network.

Identifiability analysis and Optimal Design of labeling experiments

Identifiability analysis based on the visual inspection of the response surface of the model output

Visual inspection of the response surface of the model output analysis was performed as described in the methodology section, that is, simulating the mass isotopomers of 3HA and 3HAA for different values of the metabolic fluxes. The oxidative flux of the PP pathway was set to zero. To simplify the visual analysis, all exchange fluxes of PP pathway were assumed to have the same value ($PP_{xch_1} = PP_{xch_2} = PP_{xch_3}$). The value of these fluxes is represented by the extent of reversible reaction in the figures. Including the exchange fluxes in the model can lead to identifiability problems (WIECHERT; GRAAF, 1997). Furthermore, direct measurements of these parameters are not available. Wiechert (2007) has investigated the relation of exchange fluxes with thermodynamic properties of the respective reactions. He concluded that in a situation where no net flux is associated with the reversible reaction, no physical meaning could be associated with these parameters. Despite the difficulty that the exchange fluxes bring to the problem, they must be included to complete a broad system representation. Follstad and Stephanopoulos (1997) demonstrated by simulations of carbon labeling experiments on the PP pathway that the value of the exchange fluxes can have a

considerable impact on the metabolites labeling. Follstad and Stephanopoulos (1997) also pointed out that excluding these fluxes can lead to inconsistencies between the predicted carbon label distributions using these models and those determined experimentally.

The first analysis was performed considering a medium with a mixture of 20% [U-^{13}C] glucose and 80% of natural glucose (Figure 4). The format of the response surface indicates that the flux through ED has an effect on the label enrichment only for high values of exchange fluxes. If only fully labeled G3P or natural G3P are generated, then the ED cycle has no influence in the measurements. However, when the flux through the PP pathway is high, new patterns of F6P come out, giving rise to G3P or Pyruvate presenting labeled and unlabeled carbons, and higher or lower cyclicity of the ED cycle has a significant effect.

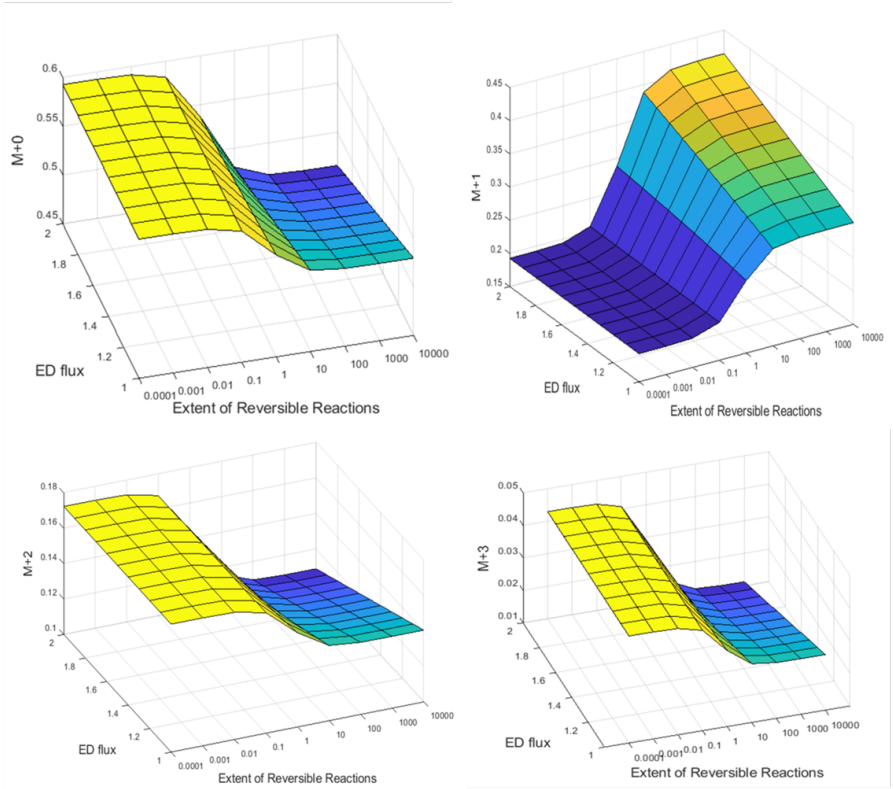


Figure 4 – Mass isotopomers simulated for different values of ED flux and extent of reversible reactions in non-oxidative PP pathway. Medium with a mixture of 20% [U-^{13}C] glucose and 80% of natural glucose.

The same global sensitivity analysis as presented above was performed for the [6-^{13}C] glucose substrate (Figure 5). The experimental design provides useful labeling information for the estimation of the ED flux in a different region of the reversible reactions that corresponds to the lower values. The data are complementary for the identification of the fluxes. As it can be observed, the amount of M+0 isotopomer increases as the flux

through ED increases as a result of the increase of cyclicity. This happens because the main effect of the cycle is in changing the labeling pattern of the third carbon atom in G3P to the first atom of F6P, which is subsequently eliminated in the form of CO_2. When PP has a high exchange flux, close to equilibrium, the effect of ED cycle is neutralized. Thus, no sensitivity information is provided in this scenario.

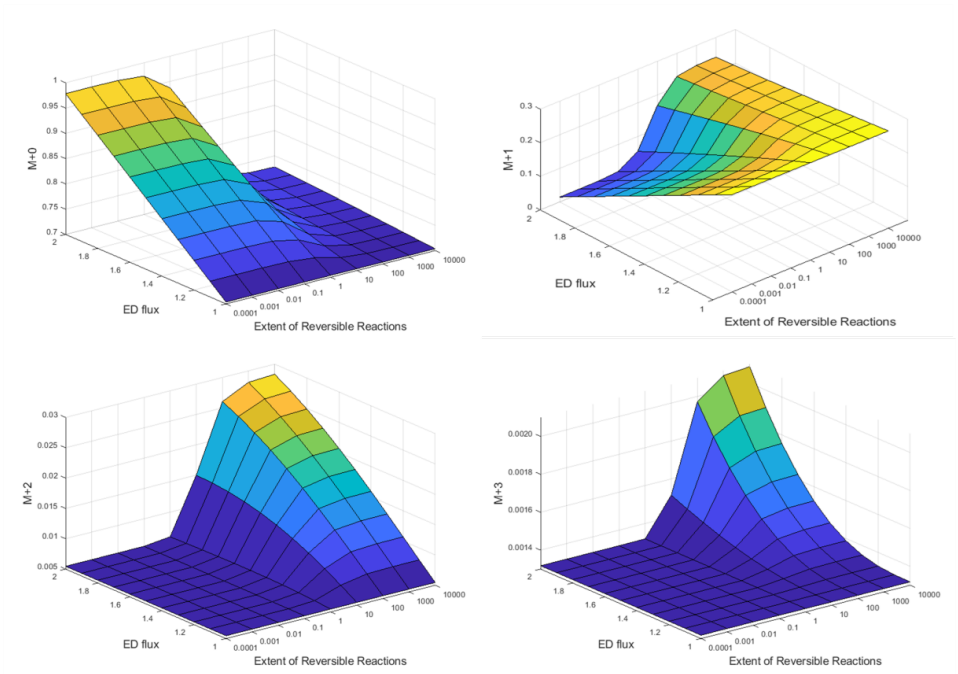
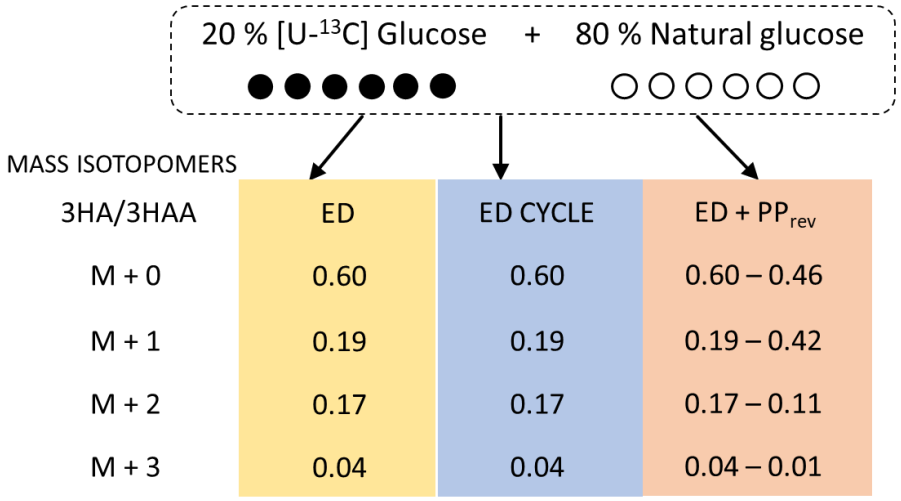


Figure 5 – Mass isotopomers simulated for different values of ED flux and extent of reversible reactions in non-oxidative PP pathway. Medium with a mixture of 55% [6-^{13}C] glucose and 45% of natural glucose.

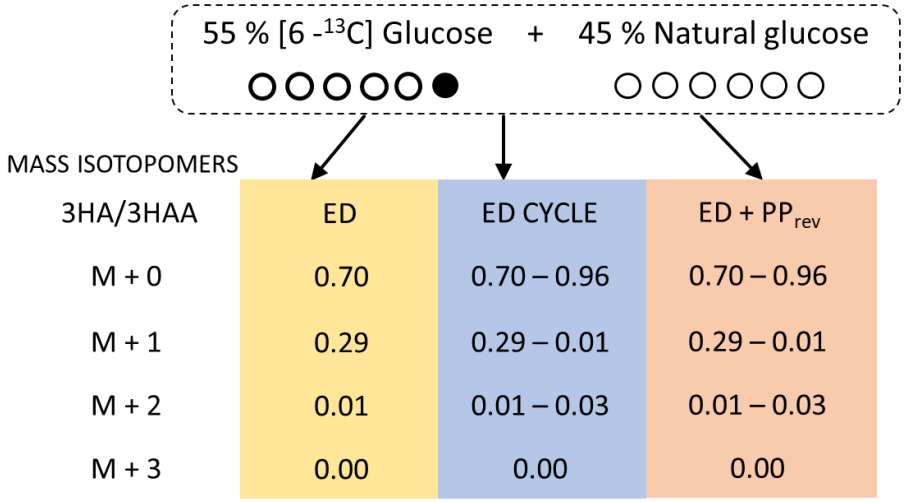
Figure 6 summarizes the results of the design of labeling experiments and sensitivity analysis. First, the simulations indicated that for experiments with [U-^{13}C] glucose (Figure 6A), the labeling pattern that will be generated by linear ED or cyclic ED is the same. However, the non-oxidative fluxes on PP pathway can shuffle the labeling, this is the main justification for this experiment. In the second case (Figure 6B), the experiments with [6-^{13}C] glucose can clearly help to distinguish from the linear to the cyclic mode operation of ED pathway. Furthermore, the effect of the fluxes in the non-oxidative PP pathway is minimal, mainly when the ED cycles operate in linear mode.

A



3HA/3HAA	ED	ED CYCLE	ED + PP_{rev}
M + 0	0.60	0.60	0.60 – 0.46
M + 1	0.19	0.19	0.19 – 0.42
M + 2	0.17	0.17	0.17 – 0.11
M + 3	0.04	0.04	0.04 – 0.01

B



3HA/3HAA	ED	ED CYCLE	ED + PP_{rev}
M + 0	0.70	0.70 – 0.96	0.70 – 0.96
M + 1	0.29	0.29 – 0.01	0.29 – 0.01
M + 2	0.01	0.01 – 0.03	0.01 – 0.03
M + 3	0.00	0.00	0.00

Figure 6 – Overview of the effect of each pathway on the labeling pattern of 3HA/3HAA mass isotopomers on experiments with [U-^{13}C] glucose and [6-^{13}C] glucose. ED – linear operation of Entner-Doudoroff pathway; ED cycle – cyclic operation of Entner-Doudoroff pathway; PPrev – non-oxidative reversible reactions of Pentose Phosphate pathway.

Fisher Information Matrix based identifiability analysis

The Optimal Design of Labeling Experiments for the study of the metabolic network of *Pseudomonas aeruginosa* LFM634 was carried out. Metabolic fluxes and the variance of the measurements were both assumed from previous experiments (OLIVEIRA, 2018). In order to restrict the labeling pattern possibilities in the glucose molecule, it was decided to

evaluate the D-criterion for glucose molecules labeled in only one of its atoms. As it can be seen in Table 1, labels on atoms 1, 2, 3 and 5 generate measurements that are not sensitive to changes in the cycle flux (ED_{flux} in Figure 4); hence that flux cannot be estimated when making experiments with these substrates. Physically, this result indicates that these atoms are eliminated in the form of carbon dioxide during the conversion of G3P into pyruvate, or that the atom does not pass through the cycle. The other two label possibilities can be used. However, the use of [6-^{13}C]glucose provides a larger value of the D-criterion because of a greater scrambling in the reactions of the PP Pathway. Therefore, this substrate was selected.

Table 1 – D-criterion for different labeled glucose substrates.

	1-^{13}C	2-^{13}C	3-^{13}C	4-^{13}C	5-^{13}C	6-^{13}C
$\sqrt[10]{D}$	0	0	0	1.0	0	2.5

In order to reduce the cost of the experiments to be performed, mixtures between [6-^{13}C]-glucose, [U-^{13}C]-glucose (fully labeled) and [U-^{12}C]-glucose (natural) substrates were evaluated using the D-criterion. Taking into account the sensitivity of the parameters with respect to changes in the mixture composition, a series of simulations were performed in a range of 0.1 between the different fractions of substrates. According to Figure 7, the experiments that provide the most information are, as expected, those with a high fraction of [6-^{13}C]glucose, the dark red region in the graph. On the other hand, the use of [U-^{13}C]glucose or [U-^{12}C]glucose as the only substrate, represented by the dark blue regions in the graph, does not allow the estimation of the desired parameter. In those conditions the problem becomes unidentifiable. The best results can be obtained for mixtures with at least 50% [6-^{13}C]glucose. Moreover, adding [U-^{13}C]glucose or [U-^{12}C]glucose to the substrate mixture has the same effect on the D-criterion, as it can be understood from the symmetry of the graph. Therefore, [U-^{12}C]-glucose was chosen because of its lower cost.

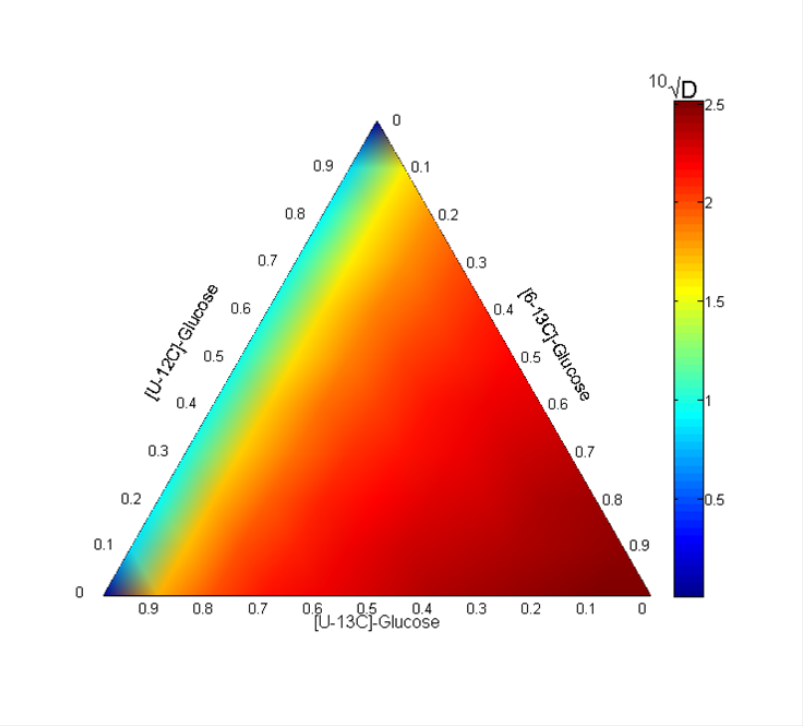


Figure 7 – D-criterion for different mixtures of substrates [6-^{13}C] glucose, [U-^{13}C] glucose and [U-^{12}C] glucose.

Principal component based identifiability analysis

Riascos et al. (2013) estimated the metabolic fluxes ratios on *P. aeruginosa* using data from an experiment with 80% of [U-^{13}C]glucose and made the hypotheses that the ED pathway operates only in linear mode and the PP pathway operates without reversible reactions. However, those hypotheses may not necessarily valid and, in this case, these pathways would shuffle the labeling of PHA mass isotopomers (KOHLSTEDT; WITTMANN, 2019). In this section, the effect of considering the aforementioned pathways in the fluxes ratio estimation problem is evaluated by an identifiability analysis.

The results presented in this section are based on a flux distribution similar to the one used by (RIASCOS et al., 2013), in which part of the flux through the ED pathway was directed to its cyclic mode (Appendix: Figure 32). PCs associated with small eigenvalues implies in a large variance of the estimates. Based on the standard deviations of measurements presented by (RIASCOS et al., 2013), eigenvalues lower than 10^{-4} were considered small (VAJDA et al., 1989). The eigenvectors (v) and eigenvalues (σ) of the Hessian matrix for the [U-^{13}C]glucose experiment is presented in Table 2. Only one of the eigenvalues is above the threshold, indicating that the problem has identifiability issues with dependencies between

parameters. The only component that can be estimated with a certain confidence is the flux through the cyclic mode of the ED. [6-^{13}C]glucose substrate was also evaluated and presented better results regarding parameter identification. As it can be seen in Table 2, only two components are below the threshold. Although the problem still presents identifiability issues, more parameters can be estimated with higher confidence with this experiment. Discovering a physically meaningful combination of the parameters from the data more precisely is not always a simple task (VAJDA et al., 1989). For this reason, the sparse eigenvectors of the hessian matrix were calculated.

Table 2 – Eigenvectors (PC) and corresponding eigenvalues of the Hessian matrix

	[U-^{13}C]Glucose				
	v_1	v_2	v_3	v_4	v_5
PP_{oxi}	-0.19	-0.06	0.08	-0.97	-0.14
ED_{cyc}	-0.92	0.36	0.04	0.16	0.01
PP_{xch_1}	-0.08	-0.33	0.83	0.04	0.44
PP_{xch_2}	-0.08	-0.29	0.33	0.19	-0.87
PP_{xch_3}	-0.33	-0.82	-0.45	0.06	0.15
σ^2	0.17	8e-18	9e-19	5e-19	2e-20
	[6-^{13}C]Glucose				
	v_1	v_2	v_3	v_4	v_5
PP_{oxi}	-0.07	0.35	0.29	-0.69	0.56
ED_{cyc}	0.37	-0.86	0.16	-0.24	0.20
PP_{xch_1}	0.59	0.21	-0.32	0.39	0.59
PP_{xch_2}	0.59	0.21	-0.32	-0.5	-0.50
PP_{xch_3}	0.41	0.22	0.82	0.25	-0.20
σ^2	60.7	0.69	0.004	9e-16	1e-16

The sparse PCs were obtained using the methodology proposed by Nakama et al. (2020). The components v_4 and v_5 were removed and the other components were sparsified, keeping the orthogonality to v_4 and v_5. The sparse PCs for [6-^{13}C]glucose experiment using the complete network are presented in Table 3. Now some combinations of parameters can be easily identified; by analyzing component one, it is clear that the objective function only depends on the PP_{xch_1}/PP_{xch_2} combination, not on these fluxes separately. Component v_2 indicates that flux ED_{cyc} can be estimated individually. A similar analysis can be applied to component v_3 and flux PP_{xch_3}. Flux PP_{oxi} cannot be estimated from this experiment since the flux is present in both components v_2 and v_3.

The metabolic network for *P. aeruginosa* can differ among strains, hence two other typical networks were also considered, the $\Delta\ ED_{cyc}$ and the $\Delta\ PP_{oxi}$ strains. For the former,

in which the ED pathway cannot operate in cyclic mode (Table 3), PP_{oxi} could be estimated separately, which was the original goal of Riascos et al. (2013). The relationship among the exchange fluxes of the PP pathways practically remains unchanged; however, component v_1 now has no influence from PP_{xch_3}. The other common variation of the *P. aeruginosa* network considered is the absence of the oxidative branch of the PP pathway (Table 3). In this case, the situation is similar to the complete network without the PP_{oxi} influence on the components.

Table 3 – Sparse PC of the Hessian matrix for the [6-^{13}C]glucose experiment.

	complete network			$\Delta\ ED_{cyc}$			$\Delta\ PP_{oxi}$		
	v_1	v_2	v_3	v_1	v_2	v_3	v_1	v_2	v_3
PP_{oxi}	0.00	-0.33	0.34	0.00	1.00	0.00	x	x	x
ED_{cyc}	0.00	0.94	0.00	x	x	x	0.00	1.00	0.00
PP_{xch_1}	-0.67	0.00	0.00	-0.71	0.00	0.00	-0.71	0.00	0.00
PP_{xch_2}	-0.67	0.00	0.00	-0.71	0.00	0.00	-0.71	0.00	0.00
PP_{xch_3}	-0.29	0.00	0.94	0.00	0.00	1.00	0.00	0.00	1.00

^{13}C-MFA of *Pseudomonas aeruginosa* LFM634

The identifiability analyses presented in the last sections contributed to a better understanding of how to estimate metabolic fluxes in *Pseudomonas aeruginosa* under non-growing conditions by our research group. All these analysis were applied to design carbon labeling experiments for the *Pseudomonas aeruginosa* LFM634 strain producing simultaneously PHA and RHL.

Glucose catabolism in *P. aeruginosa* LFM634 is exclusively by Entner-Doudoroff pathway. The Embden-Meyerhof-Parnas glycolysis is not complete in *P. aeruginosa* because it lacks the 6-phosphofructokinase enzyme (LESSIE; PHIBBS, 1984). The PP Pathway, as discussed above, presents only the non-oxidative phase and thus it has a null net flux under non-growing conditions. In spite of the fact that the ED pathway is the unique option for glucose catabolism, this pathway can operate in two distinct manners (LESSIE; PHIBBS, 1984). Glucose is converted into pyruvate and G3P, and then G3P can be converted into pyruvate (linear mode) or converted back to G6P (cyclic mode). The switch of operation modes in the ED pathway has huge consequences for the cell (SáNCHEZ-PASCUALA et al., 2019). Therefore, distinguishing between these two architectures can give insights about *P. aeruginosa* LFM634 metabolism under non-growing conditions.

A dual estimation process of metabolic fluxes was performed with data from labelling experiments with 80% (w/w) of [U-^{13}C]Glucose (Appendix: Table 15) and 55% (w/w) of [6-^{13}C]Glucose (Appendix: Table 14). Although only the data from the experiment with [6-^{13}C]glucose is sensitive to variations of the ED cycle flux, the [U-^{13}C]glucose data was needed in order to fit the PP fluxes. When data from [U-^{13}C]glucose was not used, the optimization method was unable to converge. The estimation process started from several different initial guesses, and all of them converged to the same ED flux value. The results presented a large residual for the mass isotopomers from the [U-^{13}C]glucose data (Table 4), this is an expected result because the fit for this experiment is more sensitive to the unidentifiable parameters (in practice) of the PP pathway. The simulated mass isotopomers describe better the data from [6-^{13}C]glucose experiments, thus the recycle flux ratio could be estimated with better accuracy.

Table 4 – Comparison of experimental and simulated mass isotopomers.

Mass Isotopomers	Experimental	Simulated	RMSE
M+0 [6-^{13}C]	0.8392	0.8432	0.0104
M+1 [6-^{13}C]	0.1460	0.1500	0.0127
M+2 [6-^{13}C]	0.0130	0.0055	0.0128
M+3 [6-^{13}C]	0.0018	0.0013	0.0014
M+0 [U-^{13}C]	0.6282	0.5941	0.0491
M+1 [U-^{13}C]	0.2086	0.1997	0.0132
M+2 [U-^{13}C]	0.1347	0.1658	0.0443
M+3 [U-^{13}C]	0.0285	0.0404	0.0170
Residual sum of squares		0.0024	-

The flux ratios at the estimated point (Figure 8) show a low flux in the pentose phosphate pathway. The flux in the ED cycle showed that about two thirds of G3P is recycled through G6P. This ratio is higher than the one that was recently estimated in the growth phase for *P. aeruginosa* PA01 and *P.putida* KT2440 (KOHLSTEDT; WITTMANN, 2019) that was about half of G3P. This indicates a possible increase of the cyclicity in the stationary phase due to increased NADPH demand for biosynthesis.

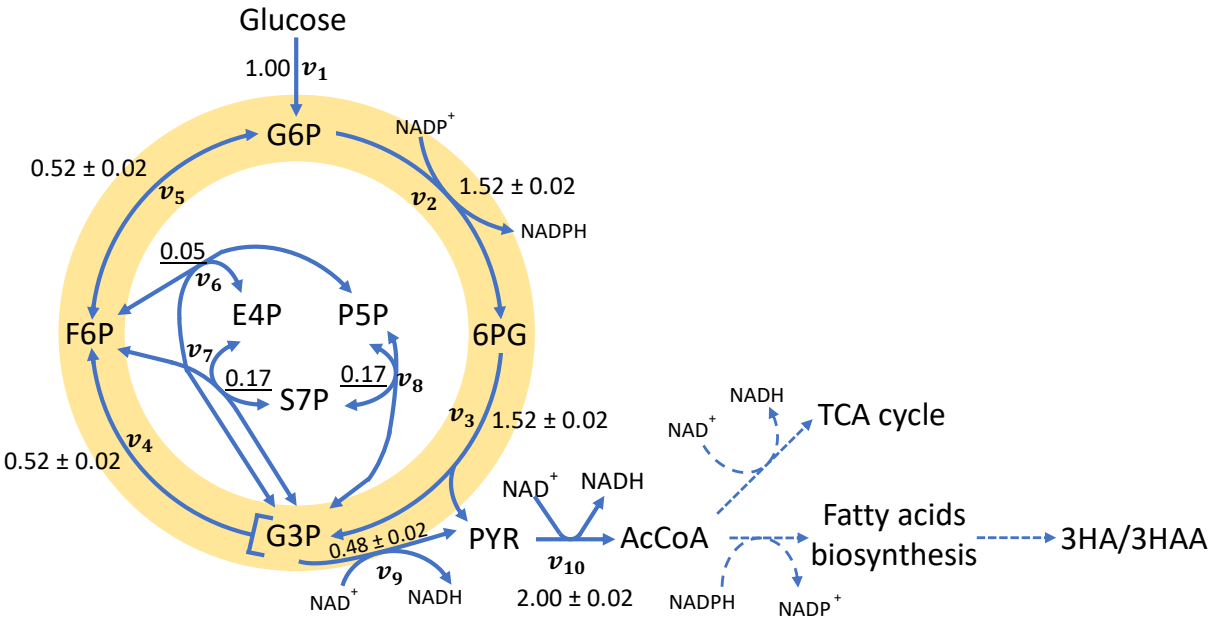


Figure 8 – Metabolic network of Pseudomonas aeruginosa LFM634 and estimated metabolic flux ratios by ^{13}C-MFA using the data from Table 4. Confidence intervals (I.C.) of 95% are shown. Metabolites abbreviations: Glucose 6-phosphate (G6P); Fructose 6-phosphate (F6P); glyceraldehyde-3-P (G3P); Pyruvate (PYR); acetyl-coenzyme A (AcCoA). Reactions: glucokinase (v1); ED cycle pathway (v_2, v_3, v_4 and v_5); PP pathway (v_6, v_7 and v_8); Pyruvate formation (v_9); Pyruvate dehydrogenase (v_10). The underlined fluxes correspond to exchange fluxes.

Glucose metabolism in *P. aeruginosa* is cyclic when producing PHA and RHL

The cyclic operation of the ED pathway seems to play a key role in NADPH reduction power regulation in *Pseudomonas*. This architecture was also observed in several others microorganisms (PORTAIS; DELORT, 2002). In *Pseudomonas*, it was observed in cells subject to environmental stress (NIKEL et al., 2015), and in alginate biosyhthesis (LYNN; SOKATCH, 1984). The results suggests that the cyclic mode plays a key role in the simultaneous production of PHA and rhamnolipid. Biosynthesis for production of PHA and rhamnolipid has a higher demand for NADPH supply than the growth phase (NIKEL et al., 2015; KOHLSTEDT; WITTMANN, 2019), which, in combination with precursors like Acetyl-CoA, are the building blocks of those biocompounds. This reinforces the fact that the NADPH/AcCoA flux ratio is relevant for the understanding of the biosynthesis metabolism.

The higher the cyclicity in the ED pathway higher the NADPH/AcCoA flux ratio reached (Figure 9). Previous studies have also shown a periplasmic oxidation of glucose into gluconate in *Pseudomonas* (NIKEL et al., 2015; KOHLSTEDT; WITTMANN, 2019), however, the

periplasmic route does not shuffle the carbon atoms, then it is not possible to estimate this flux by ^{13}C-MFA. As expected, the stoichiometric maximum of the NADPH/AcCoA flux ratio (ratio=1.00) was obtained for the total recycle of G3P in the ED pathway and null gluconate formation. Since two-thirds of G3P is recycled, a maximum NAPDH/AcCoA flux ratio of 0.76 mol/mol is expected.

The ideal NADPH/AcCoA flux ratio depends on which monomer is synthesized, for the 3HA present in higher amounts on PHA and RHL (3HD), the ratio is 1.4 mol/mol. The estimated fraction of G3P recycled was 68%, meaning a maximum NADPH/AcCoA flux ratio of 0.76 mol/mol, thus a supplementary source of NADPH must be active. Furthermore, the simulated NADPH/AcCoA ratio can be lower if cofactor specificities are considered. Although glucose-6-phosphate dehydrogenase enzyme in *P. putida* and *P. aeruginosa* was considered as having a preference for NADPH (NIKEL et al., 2015) that preference could be indeed much lower (OLAVARRIA et al., 2015; CARDINALI-REZENDE et al., 2020).

Three possible sources of redox power (NADPH) are isocitrate dehydrogenase, maleic enzyme and transhydrogenase (NIKEL et al., 2015). The two first imply in a flux through the TCA cycle, thus less AcCoA would be generated, and consequently, less PHA and rhamnolipids. The transhydrogenase enzyme would lead the cell to maximum yield, with the conversion of NADH into NADPH. NADH production is only affected by the G3P node in the central metabolism. As the recycle flux increases, less flux goes to pyruvate (glyceraldehyde-3-P dehydrogenase enzyme), therefore more NADPH is produced instead of NADH (Figure 8). The NADH/AcCoA ratio estimated here considering reactions 2 and 10 was 1.24. This value, along with the NADPH generation by the ED cycle is more than sufficient for supplying PHA biosynthesis. Therefore, future work should measure the activity of transhydrogenase enzyme for a complete scenario of the NADPH/AcCoA flux ratio under 3HA biosynthesis.

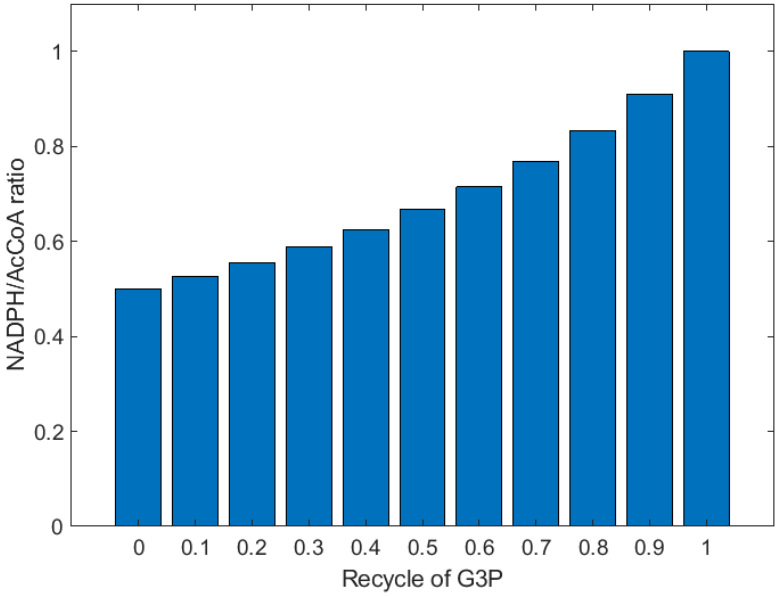


Figure 9 – NADPH/AcCoA ratio variations in relation to recycle ratio of G3P.

2.4 Conclusions

Identifiability analysis is an essential tool for research in which ^{13}C-MFA is applied. This analysis can reduce experimental costs and increase the precision of the estimated metabolic fluxes. The DOE is crucial because, as already demonstrated, the wrong choice of labeled substrate may result in misinterpretation of data (KOHLSTEDT; WITTMANN, 2019). As examples of studies using *Pseudomonas*, the ED cycle was underestimated in Berger et al. (2014), and the PP pathway overestimated in Opperman and Shachar-Hill (2016).

The identifiability analysis based on the visual inspection of the response surface of the model output was shown to be useful for obtaining a global picture of the effect of the fluxes on the measurements, however, it is not scalable for larger metabolic networks. The analysis based on FIM is a practical and straightforward way of performing the identifiability analysis, but an *a priori* flux distribution must be provided. Furthermore, it was demonstrated that the application of a sparse PC decomposition can improve the identification of dependencies among parameters and the determination of estimable fluxes from available data. The method presented in this work enables the identification of parameter combinations that can be estimated from labeling experiments. This represents an advantage when compared to the methods presented in literature that only verifies whether a subset of parameters is identifiable, leading to a combinatorial problem (WINDEN et al., 2001a; KAPPELMANN et al., 2016).

The methodology applied in this work provided useful information for elucidating flux problems using conventional methods of analysis. This study was useful for bringing up information about the simultaneous production of PHA and RHL by applying metabolic network analysis. The measurements of HAs and HAAs were obtained under stationary growth phase and analyzed by standard techniques like GC-MS. The results indicated that experiments with [U-^{13}C]glucose and [6-^{13}C]glucose would be adequate for the estimation of the flux ratio. The analysis allowed us to conclude that the PP pathway has a low flux and the ED pathway operates in a cyclic mode, playing a key role in the regulation of the NADPH/AcCoA flux ratio, which is crucial for PHA and 3HAA biosynthesis. The data revealed in this work will be useful to establish better metabolic models for *P. aeruginosa* LFM634. *In silico* experiments will be used to identify different approaches (genetic modification, culture conditions etc) aiming to improve the production of PHA and/or RHL as well as to evaluate this biotechnological chassis for the synthesis of other bioproducts.

3 Surrogate model approximation of Flux Balance Analysis

In recent decades, bioprocessing technology has emerged as a sustainable alternative to produce a wide range of chemical products. However, key limitations of bio manufacturing systems, such as high production costs and feedstock variability, still impose a burden when competing with traditional processes (e.g. petrochemical) (BURG et al., 2016). For addressing these shortcomings and ensuring a cost-effective production, advanced process control techniques, like Economic model predictive control (EMPC), can be applied. They have a proven potential to maximize profit, increase process regularity, and enhance product quality (LEE, 2011).

In EMPC, a dynamic model is used for predicting the system response to changes in key process variables. Based on this model, the controller decides the best trajectories of these variables for minimizing/maximizing a given criterion, while satisfying a set of operational constraints. For example, EMPC finds the best possible feeding strategy of a fed-batch reactor that minimizes the difference between the end-batch concentration and a given desired value, while guaranteeing that the reactor temperature stays below a threshold. Despite being widely used in other industries, EMPC's full potential in relation to biological processes is yet to be achieved (MEARS et al., 2017). The main drawback is linked to the dynamic models used in the controller. EMPC requires models that can be reliably solved in different operating conditions but are also sufficiently detailed to describe the real process (SOMMEREGGER et al., 2017).

Obtaining such dynamic models for controlling systems where the microorganisms metabolism needs to be taken into account becomes a critical issue. Models that do not include intracellular metabolism (unstructured models) are the most common in the field, but they do not describe crucial features of the metabolism (HODGSON et al., 2004). On the other hand, dynamic flux balance analysis (dFBA) models can be applied for accurately describing the microorganisms, but are not appropriate for computationally intense optimization studies (NIELSEN, 2017). These models consist of mass balances for extracellular metabolites, kinetic relations for nutrients uptake, and Flux Balance Analysis (FBA) for intracellular metabolites, which is performed by an optimization scheme that determines the metabolic fluxes (MAHADEVAN et al., 2002). Therefore, solving it inside the EMPC culminates in a bi-level optimization problem that is a challenging task to be reliably carried out online.

In this work, we propose to use a dFBA model together with a model predictive control in order to maximize a desired bioproduct formation in a fed-batch bioreactor. However, instead of solving the controller problem as a bi-level optimization, we replace the FBA problem by a surrogate model. For obtaining this model, the linear programming problem of FBA is solved offline for a wide range of feed conditions, here represented by the uptake rate. Next, a great number of simulation points is obtained on a grid covering the whole operating space for each exchange flux of extracellular metabolites in the model. Then, a polynomial model is fitted by Partial Least Square Regression (PLSR) in order to approximate the behaviour of each exchange flux. The resulting model is then applied in the controller. This novel approach allows us to take advantage of the prediction capabilities and accuracy of the dFBA models, while guaranteeing that the controller problem can be reliably solved online using standard optimization tools. Moreover, the methodology opens the possibility for applications of genome-scale models of metabolism on macromolecular expression models (LLOYD et al., 2018), dynamic regulation of metabolism (ANESIADIS et al., 2008) and process simulators, like ASPEN PLUS (PLUS, 2003). A diagram of our approach is shown in Figure 10.

In order to validate the approach, a case study of a fed-batch bioreactor of *Saccharomyces cerevisiae* is used. The controller objective is to maximize ethanol production. After an offline identification step, we obtain a surrogate model using PLSR and couple it to the EMPC. The results show that the control strategy provides a higher ethanol titer in comparison to the open-loop operation. Additionally, since the surrogate approximation of FBA allows us to apply the more detailed *Saccharomyces cerevisiae* reconstructed genome-scale network model available up to date (see Lu et al. (2019)), the overall model presents a good accuracy and a similar prediction capacity from models that solve FBA using an optimization scheme.

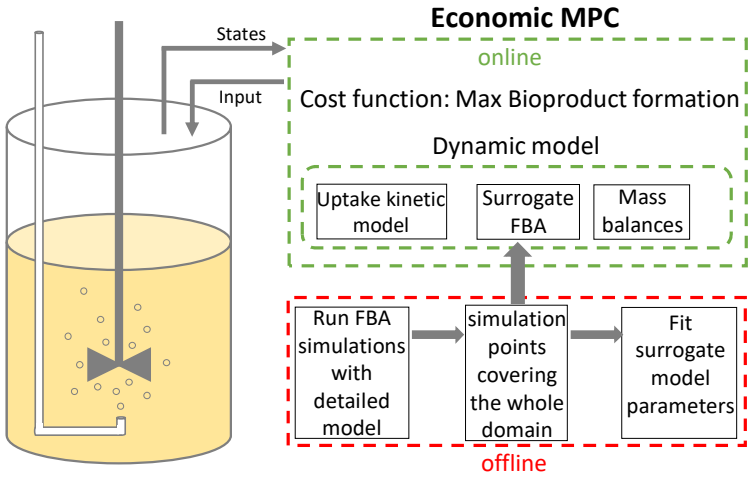


Figure 10 – Economic MPC of a bioreactor using surrogate FBA model scheme.

Remark. *The results presented here are based on the works Oliveira et al. (2021a) and Oliveira et al. (2022). Originally the surrogate FBA methodology was developed for application of dFBA models in Economic MPC of bioreactors. The surrogate FBA was also extended to the parameter estimation of dFBA kinetic parameters. This chapter focus is the formulation of EMPC with surrogate FBA as described in Oliveira et al. (2021a) (Study case 1). Also, the parameter estimation application was included as described in Oliveira et al. (2022) as Study case 2.*

3.1 Research background

Flux Balance Analysis

Flux balance analysis (FBA) has become the most popular mathematical method for simulating metabolism using GSM in the past years (MARANAS, 2016), and the reason for that relies on the simplicity and applicability of the method. FBA is a method based on an optimization approach using the stoichiometric matrix. For all the possible fluxes for each reaction in a network (Figure 11), the stoichiometry of the network imposes some constraints for the possible solutions. FBA uses some criterion, i.e. objective function, to select one phenotype among all possibilities. Mathematically, an FBA can be formulated as a linear programming problem:

$$\begin{aligned} \max \quad & c^T v \\ \text{subject to:} \quad & S \cdot v = 0 \\ & lb \leq v \leq ub \end{aligned} \tag{3.1}$$

where lb and ub are the lower and upper bound vectors for metabolic fluxes (v) respectively. S is the stoichiometric matrix. c is the vector of coefficients that multiplies the flux vector in order to express the objective function. The most applied objective function on FBA is to maximize biomass formation (MARANAS, 2016), where vector c is a vector with zeros except for the biomass reaction. The fluxes constraints are typically measured external fluxes or derived from thermodynamics data. The *ad hoc* biomass reaction is built in the reconstruction process of the network, and it is a drain reaction of all biomass precursors (e.g., amino acids, lipids, and carbohydrates). The fluxes in FBA usually have unity of $\frac{mmol}{gDWh}$ where gDW means grams of dry cell weight. The biomass reaction has unity of h^{-1} and to build the biomass reaction, the chemical composition of the cell must be determined experimentally.

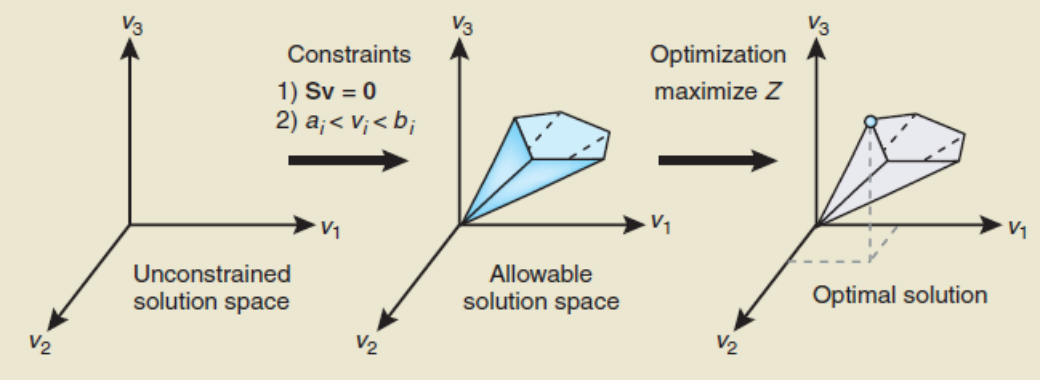


Figure 11 – The conceptual basis of constraint-based modeling.

Source: (ORTH et al., 2010)

Dynamic Flux Balance Analysis

The dynamic version of the FBA model, the dFBA, is formulated assuming that the intracellular dynamics metabolism is much slower than the extracellular dynamics. Thus, the internal metabolism can be represented by the FBA model, while the extracellular metabolites are modeled by mass balance equations. Also, kinetic expressions are used to model the uptake of substrates. dFBA can be formulated as an ordinary differential equation system with an embedded optimization (ODEO) problem:

$$\begin{aligned} \frac{dx}{dt} &= F(x, v) \\ v_{uptake} &= g(x) \\ x &\geq 0 \\ v &= FBA(v_{uptake}) \quad \text{(Equation 3.1)} \end{aligned} \tag{3.2}$$

where x is the concentrations of the external metabolites; F is the right-hand side of the mass balance equations; and g is the kinetic expression to calculate the uptake fluxes v_{uptake}. The dFBA model is an alternative to the utilization of kinetic models of metabolism that need the knowledge of enzymatic reactions mechanism and kinetic parameters from the internal reactions. The dFBA only uses a few kinetic parameters on the uptake reactions that are considered as a limiting step. However, despite being easy to build, dFBA models as an ODEO problem are challenging to solve. The most applied methods to solve dFBA models are (GOMEZ et al., 2014):

1. **Static optimization approach (SOA)**: For each integration interval, the uptake rates are calculated from kinetic expressions and used as constraints in the FBA optimization problem. The optimal extracellular fluxes $\boldsymbol{v}^{\star}$ are, then, fed to the dynamic model integrator. Since these fluxes are assumed constant during the entire integration interval, they can be directly included in the differential equations. Next, the system of equations is integrated in time for computing the state variables and uptake rates. This process is repeated until the final simulation time;
2. **Dynamic optimization approach (DOA)**: The FBA is optimized over the entire time period applying an instantaneous objective function. The problems is solved as a nonlinear programming problem (NLP) by converting the set of ordinary differential equations into an algebraic equation system by orthogonal collocation on finite elements;
3. **Direct approach (DA):** The DA solves the Linear Problem (LP) of FBA inside the ordinary differential system of equations.

Each of these approaches has its advantages and disadvantages. The SOA approach, for example, has a trade-off between performance and computational time because small integrator steps are usually required for convergence (GOMEZ et al., 2014). On the other hand, the DOA approach can take advantage of robust nonlinear solvers (WÄCHTER; BIEGLER,

2006), but it is limited to small-scale metabolic models because of the significant dimension increase due to time discretization (GOMEZ et al., 2014). Finally, the DA method needs the solution of an LP (FBA) problem in each evaluation of the ODE system. Hence, the DA method can be time consuming. Implicit ODE integrators with adaptive step size for error control can be applied to reduce the number of steps sizes (GOMEZ et al., 2014).

Even though each of the methods listed above presents an specific strategy to solve the dFBA problem, the surrogate model approximation of the FBA can be coupled to all of them, which is an important benefit of this approach.

Model Predictive Controller

Model predictive controller is used for determining the values of the system manipulated variables that minimize/maximize an objective function. Typically, this objective function is selected such that the controller regulates the system outputs at previously determined setpoints (e.g. maintaining the outlet concentrations of a reactor at a given level). If the MPC is used for solving a production optimization problem, in which the objective function reflects some economic-like criterion, it is referred to as Economic Model Predictive Control (EMPC)(RAWLINGS et al., 2017). The EMPC problem can be written as:

$$
\begin{aligned}
&\min_{\mathbf{u}} \quad \int_{t_0}^{t_0+T_p} \phi(\mathbf{x}(t), \mathbf{u}(t))dt \\
&\text{s.t.} \\
&\mathbf{x}(0) = \mathbf{x}_0 \\
&\dot{\mathbf{x}}(t) = \mathbf{f}(\mathbf{x}(t), \mathbf{u}(t)) \qquad t \in [t_0, t_0 + T_p] \\
&\mathbf{y}(t) = \mathbf{h}(\mathbf{x}(t)) \qquad t \in [t_0, t_0 + T_p] \\
&\mathbf{y}_{\min} \leq \mathbf{y}(t) \leq \mathbf{y}_{\max} \qquad t \in [t_0, t_0 + T_p] \\
&\mathbf{u}_{\min} \leq \mathbf{u}(t) \leq \mathbf{u}_{\max} \qquad t \in [t_0, t_0 + T_p]
\end{aligned}
\tag{3.3}
$$

where, $\mathbf{x}$ are the system states, $\mathbf{u}$ the system manipulated variables, and $\mathbf{y}$ the process outputs (measurements). The inputs and outputs bounds are $\mathbf{u}_{\min}$, $\mathbf{u}_{\max}$, $\mathbf{y}_{\min}$ and $\mathbf{y}_{\max}$, respectively. $\mathbf{f}(\cdot,\cdot)$ is the system dynamic model and $\mathbf{h}(\cdot)$ is a function that maps the states to the system outputs. $\phi(\cdot,\cdot)$ is the controller economic objective function. The current time step is represented by t_0, and T_p represent the controller prediction horizon.

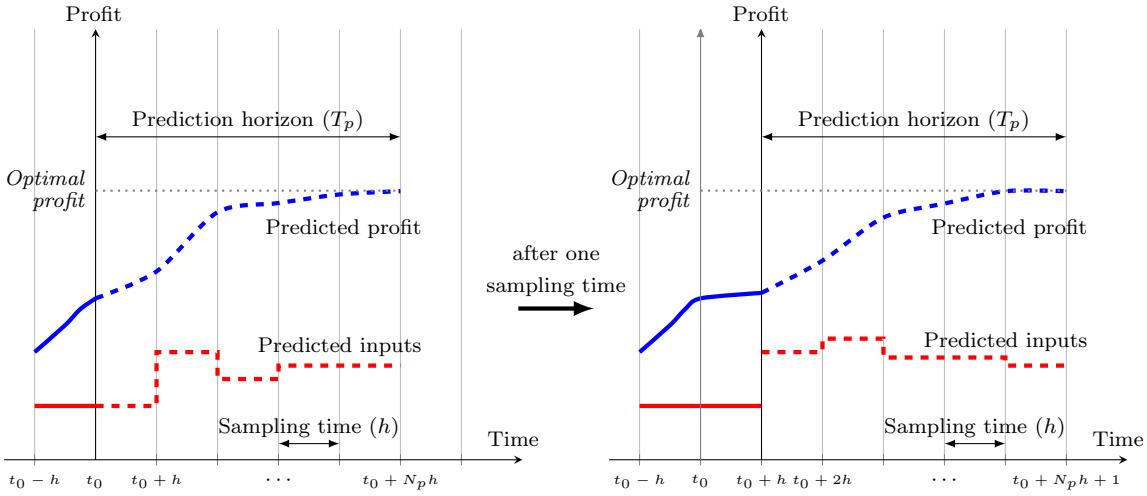


Figure 12 – Simplified diagram of an economic model predictive controller.

On an EMPC implementation, the objective is to maximize a profit function. Figure 12 shows a simplified diagram of an EMPC implementation, in which the objective is to maximize a profit function. At each time step, the current plant information is fed to the controller. Then, the EMPC defines the optimal input trajectory (red dashed line) for maximizing the system profit (blue dashed line). For practical reasons, the prediction horizon is truncated. It is defined as a multiple of the controller sampling time h, i.e. $T_p = N_p \cdot h$.

Although a sequence of inputs $\mathbf{u}^\star = [\mathbf{u}_1^\star, \mathbf{u}_2^\star, \ldots, \mathbf{u}_{N_p}^\star]$ is computed, only $\mathbf{u}_1^\star$ is implemented. Then, after one sampling time, the controller receives the new plant information and a new input sequence $\mathbf{u}$ is computed using Equation (3.3). The control problem is solved in this receding horizon framework because the $\mathbf{u}^\star$ trajectory may change due to unexpected system disturbances. Additionally, the model may not be a perfect representation of the system and this feedback strategy mitigates the plant-model mismatch (RAWLINGS et al., 2017).

Since the controller problem needs to be solved *every* sampling time, using a system model $f(\cdot, \cdot)$ based on dFBA can be challenging for practical reasons. For obtaining the solution of Equation (3.3), the system dynamic model needs to be integrated in time. If integration solvers are applied, the FBA optimization problem needs to be solved at every step of the solver. Event-based algorithms that detect LP base changes are state of the art for dFBA simulations(GOMEZ et al., 2014). The points where the events take place are non-differentiable points. Hence, the lack of model gradients hinders the application of

strategies that simultaneously integrate the system and solve the optimization problem, like collocation-based methods. Alternatively, one can use statistical/random methods, like genetic algorithms. However, the high dimensionality of the flux space (e.g. over 3900 reactions in *Saccharomyces cerevisiae* metabolic network) makes the problem intractable since these methods do not scale very well for large systems. Also, there is no rigorous convergence guarantee to the problem solution when statistical/random methods are used. Therefore, a more flexible and reliable modeling strategy for the dFBA needs to be adopted, which is the main contribution of this work.

Surrogate model approximation of FBA

In order to obtain a more flexible and reliable model for the controller, the LP (FBA) problem can be replaced by a surrogate model. This meta-modeling strategy has been applied for optimizing a wide range of processes (e.g. oil gas field production (GRIMSTAD et al., 2016) and pharma manufacturing (ÖNER et al., 2020)). Recently, a surrogate model has been applied to uncertainty quantification on dFBA models (PAULSON et al., 2019). The goal is to represent non-linear complex systems by functional approximations, such as polynomials, that are computationally cheaper to evaluate and can easily provide gradient information (QUEIPO et al., 2005). Despite being the simplest choice, using polynomial models to represent the dFBA non-smooth behavior(MAHADEVAN et al., 2002) can be challenging. One alternative is then to divide the domain into subdomains, in which the behaviour is smooth. Then, a polynomial model is fitted to each of the subdomains.

After defining the surrogate model structure, we perform a series of simulations using the full genome-scale network for obtaining a data collection that will be used to estimate the parameters of the surrogate model. The goal here is to map the uptake fluxes to extracellular and biomass fluxes, and replace the solution of the LP (FBA) optimization problem.

Before running the simulations, we determine an equidistant square grid for the independent uptake fluxes variables. Since this step is carried out offline, the grid can be chosen as fine as necessary for describing the metabolic network at hand. The grid shape can also be adjusted according to the system of interest. As is well known for larger problems the equidistant square grid can suffer from the so-called curse of dimensionality. Fortunately, new methodologies for the sampling method step have been proposed in literature. Sparse grids are a very common method to reduce the number of sampling points (PFLüGER et al.,

2010). Also, more recently, a methodology that applies PLSR as a decision criterion of the number of sample points has been proposed (STRAUS; SKOGESTAD, 2019). However, we would like to point out that the sampling for the surrogate model identification is done offline and can be easily parallelized. Thus, even if we use a equidistant grid, this step does not affect the real-time performance of the controller.

For each grid node, we run an FBA optimization problem subjected to the correspondent values of uptake fluxes. In our approach, we perform a Flux Variability Analysis (MAHADEVAN; SCHILLING, 2003) to ensure that the effect of multiple optimal solution has a minimum impact in extracellular fluxes values. For the cases where alternative optima may be relevant, it is possible to replace the FBA simulations by parsimonious FBA simulations (LEWIS et al., 2010) or performing lexicographic optimizations as described by Gomez et al. (2014).

The surrogate models are written as follows:

$$\upsilon_{out,i} = \sum_{j=1}^{n_p} \prod_{k=1}^{n_u} a_{ij} \upsilon_{in,k}^{\alpha} \tag{3.4}$$

where, $\upsilon_{out,i}$ is the approximate i^{th} output flux; $\upsilon_{in,k}$ are the uptake fluxes that range from $k = 1, \cdots, n_u$; a_{ij} are the polynomial coefficients. For each output flux i, $j = 1, \cdots, n_p$, where n_p is the number of parameters of the surrogate models; $\alpha \in [0, n_\alpha]$ are the integer exponents of the polynomial terms. The values of a_{ij} are determined using a Partial least squares estimator (PLS). This method chooses the parameters components (the polynomial and interaction terms) that present the maximum covariance between independent and predicted variables, which is useful to avoid data over-fitting (GELADI; KOWALSKI, 1986).

As an additional measure to avoid over-fitting, we divide the operation region in L different zones. Next, we fit low-order polynomials to each zone instead of using a single high-order polynomial model to describe highly nonlinear input and output fluxes (NIELSEN, 2017). The simple polynomial models are connected by a smooth analytic approximation of the heaviside function, which makes them infinitely differentiable (i.e. of class C^{∞}):

$$\upsilon_{out,i} = \sum_{l=1}^{L} \frac{\upsilon_{out,i,l}}{1 + e^{-k_l H(\upsilon_{in})}} \tag{3.5}$$

where, k_l is the heaviside parameter. $H(\upsilon_{in}) := (\upsilon_{in} - \upsilon_{lim}) \in \mathcal{R}^{n_u} \rightarrow \mathcal{R}$ is a function that maps the intake fluxes into a scalar. If $\upsilon_{in} > \upsilon_{lim}$, the l^{th} term of the sum becomes $\upsilon_{out,i,l}$, otherwise

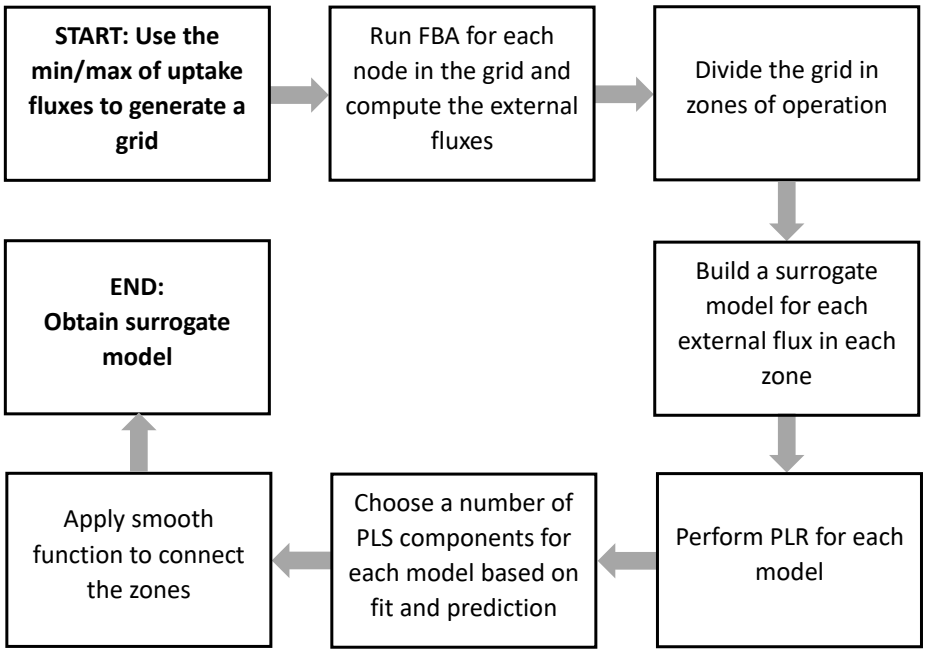


Figure 13 – Diagram block describing the surrogate model identification process.

it becomes 0. Figure 13 shows a diagram block summarizing how we obtain the surrogate models:

After obtaining the surrogate models that replace the LP (FBA) problem, Equation (3.5) is included in the controller formulation, Equation (3.3), as part of the dynamic model $f(\cdot,\cdot)$.

Nonlinear model predictive control of bioreactors using dFBA

Different bioprocess control strategies, in which a dFBA model is included in the controller, have already been reported in the literature. They are focused on simplifying/reducing the metabolic network, on reformulating the bi-level optimization problem, or on using a surrogate model.

Jabarivelisdeh et al. (2020) worked with *E.coli* with a small-scale metabolic model (50 metabolic reactions) in order to solve the bi-level optimization problem. (ZHENG et al., 2020) applied a tableau based tree method for a robust EMPC formulation, that also culminates in a bi-level problem. An *E.coli* network of four metabolic reactions was used.

The utilization of small-scale metabolic network models allows the controller problem to be solved online, but it may lead to significant plant-model mismatch. Neglected metabolic reactions can affect the model dynamic behavior and the prediction of the flux model

response to extracellular stimuli (FRITZEMEIER et al., 2017). Since the model is used for determining an input trajectory to an *a priori* unknown optimal value in EMPC, the deviation between model predictions and the system behavior can lead to a suboptimal (and even infeasible) input sequence. This can significantly affect the controller performance even with the receding horizon strategy. An additional shortcoming in this simplification strategy is the fact that the controller relies in the plant measurements, which are used to updated the model states x, to compute the next input sequence. If a measurement is missing (e.g. probe failure), the controller still needs to provide a solution in the next sampling time. In this case, the adequacy of the computed input sequence heavily depends on the controller model accuracy, which can be considerably impacted by plant-model mismatch.

A second alternative is reformulating the system. Chang et al. (2016) replaced the LP problem of the DFBA with its first optimally conditions, which are included as equality constraints in the controller formulation (Equation (3.3)). By solving this augmented system, the solution of the FBA and control problem are obtained simultaneously. The authors applied this strategy to maximize the ethanol production in a fed-batch bioreactor with *Saccharomyces cerevisiae*. Since this reformulation led to a large system of equations and variables, the authors used a small metabolic network complexity to decrease the computational burden.

The alternative of using surrogate models to replace dFBA models has also been explored by Kumar and Budman (2015) and Kumar and Budman (2017) who applied a polynomial chaos expansion for the dFBA model in the implementation of a robust EMPC for an *E. coli* batch fermentation. The method has shown promising results to deal with the parametric uncertainty. However, the resulting formulation for online application of the EMPC is still a bi-level optimization architecture. In both works, a *E.coli* network composed of four metabolic reactions was used.

Even tough the approaches above have demonstrated great potential for using metabolic network models on EMPC-based control, solving full genome-scale network online based on a surrogate model can significantly increase the model accuracy when compared to the reduction strategies listed above, whereas it keeps the model simple enough that it can be reliably solved online. Furthermore, given that the surrogate models are properly identified, we can decrease the loss of optimally associated with the use of this model approximation even without explicitly solving for the first optimally conditions of the FBA problem, like in Chang et al. (2016).

It is important to make a clear distinction between the method proposed here and the methods proposed by Kumar and Budman (2015) and Kumar and Budman (2017). While Kumar and Budman (2015) and Kumar and Budman (2017) trained a surrogate model for the entire dFBA model, here, only the FBA LP inner problem is replaced by a surrogate model. As it will be shown, this methodology allows that more complex metabolic-networks be used and also transforms the problem into a single level optimization.

Another possibility would be to use network reduction. Several methods have been proposed in literature in the past years (SINGH; LERCHER, 2021). Even though these methods can in some cases describe the behaviour of the original network, they can also reduce the model versatility (SINGH; LERCHER, 2021). Also, in some cases, the reduced network could yet have a prohibitive size for MPC bi-level applications.

Parameter estimation in dynamic metabolic models

Kinetic models of cell metabolism are a very promising category of models because of their high prediction capabilities. However, the lack of knowledge of enzymatic reactions mechanisms and the need to estimate a prohibitively number of parameters makes the application of kinetic models restricted to the description of small enzymatic pathways so far. Dynamic Flux Balance Analysis (dFBA) models appear as an alternative approach where the internal flux distribution is described by a steady-state model and the solution computed by an optimization problem. Therefore, in dFBA models, only a small number of parameters must be estimated from experimental data making the problem solvable in practice. Despite the fact that the number of parameters to be estimated is reduced, dFBA model consists of a system of differential equations and an embedded optimization problem, making the parameter estimation problem challenging to solve.

The sequential solution of the problem consists in solving the nested LP and the ODE system inside the optimization. The lack of gradient information and the non-smoothness of dFBA makes the problem hard to solve. Leppävuori et al. (2011) developed a sequential gradient-based solution with direct sensitivities equations. They estimated 8 parameters and used a metabolic network of 1266 enzymatic reactions and 1061 metabolites. Waldherr (2016) reformulated the bi-level problem as a Mixed Integer Quadratic Program, however, due to the computational burden they used a small-scale network of 10 reactions and 12 metabolites. Raghunathan et al. (2003) and Raghunathan et al. (2006) reformulated the

problem as a Mathematical Program with Complementary constraints (MPCC). MPCC cannot be solved by standard NLP solvers, therefore they relaxed the complementary constraints using a barrier parameter. They applied the approach to a small-scale metabolic network of 39 reactions and 43 metabolites.

Here, we investigate the suitability of the surrogate FBA methodology in order to reduce the computational load of parameter estimation problems using dFBA models. As a case study, a dFBA model is formulated to describe batch cultivation of glucose and xylose mixtures by *Saccharomyces cerevisiae*. *S. cerevisiae* is the main microorganism for industrial alcoholic fermentation; however, the spectrum of substrates is almost restricted to sugars, such as glucose and fructose. *S. cerevisiae* does not naturally consume xylose and the development of strains of *S. cerevisiae* for the conversion of xylose into ethanol by *S. cerevisiae* has been implemented (KUYPER et al., 2004). However, studies are needed in order to make the xylose fermentation by *S. cerevisiae* more efficient. Metabolic models can be very useful to achieve this aim, because they allow the understanding of xylose fermentation in a multi-scale approach, from a genome-scale level to the bioreactor operation.

3.2 Methodology

Study case 1: Economic MPC

For investigating the capabilities, advantages and challenges of the implementation of our approach, we use the fed-batch bioreactor containing *Saccharomyces cerevisiae* model proposed by Chang et al. (2016). The controller goal is to maximize ethanol production by manipulating the glucose feed and the dissolved oxygen level. The manipulation of the glucose feed profile is important in order to avoid substrate inhibition effects. In turn, the manipulation of the dissolved oxygen profile is important to optimize the switch from aerobic to anaerobic regime.

The model is based on the work of Chang et al. (2016). It is composed by component mass balances, total mass balance (since the density is assumed constant, this equation becomes a volume variation relation) and equations for computing the glucose, oxygen consumption and biomass and ethanol exchange fluxes. The direct manipulation of oxygen concentration was performed, assuming a perfect feedback controller. Therefore, oxygen

balance was not included in the model. The percent dissolved oxygen is calculated as a function of the oxygen saturation concentration ($DO = O/O_{sat}$). The model can be written as:

$$\begin{aligned}
\frac{dV}{dt} &= F \\
\frac{d(VX)}{dt} &= \mu VX \\
\frac{d(VG)}{dt} &= FG_f - v_g VX \\
\frac{d(VE)}{dt} &= v_e VX \\
v_g &\leq v_{g,max} \frac{G}{K_g + G + (G^2/K_{ig})} \frac{1}{1 + (E/K_{ie})} \\
v_o &\leq v_{o,max} \frac{O}{K_o + O} \\
X, G, E &\geq 0 \\
v_e,\ \mu &= \Xi(v_o, v_g)
\end{aligned} \tag{3.6}$$

where, X, G, O and E are the concentrations of biomass, glucose, oxygen and ethanol, respectively. F is the feed flow rate, which contains the growth rate limiting nutrient glucose with a concentration G_f, and other essential nutrients that are considered to be supplied in excess. V is the reactor liquid volume. v_g and v_o are the uptake fluxes of glucose and oxygen that are calculated based on the oxygen and glucose saturation parameters (K_o and K_g) and ethanol and glucose inhibition constants (K_{ie} and K_{ig}).

The exchange flux of ethanol v_e and relative cellular growth μ are computed by the mapping Ξ, which can be solved either by the FBA problem in Equation (3.1) or by the surrogate approximation in Equation (3.5). The metabolic network model used was the consensus yeast metabolic network (HEAVNER et al., 2013) version 8.3. The network consists of 2666 metabolites and 3928 enzymatic reactions. Table 5 lists the model parameters values, and the system concentration initial values.

Table 5 – Dynamic Flux Balance (dFBA) model parameters from Chang et al. (2016).

Parameter	Description	Value [unit]
$X(0)$	Biomass initial condition	0.10 $[g]$
$G(0)$	Glucose initial condition	7.32 $[g]$
$V(0)$	Volume initial condition	0.5 $[L]$
G_f	Feed glucose concentration	50 $[gL^{-1}]$
K_o	Oxygen saturation	3.0 x 10^{-6} [molL^{-1}]
K_g	Glucose saturation	0.5 $[gL^{-1}]$
K_{ig}	Inhibition constant	10 $[gL^{-1}]$
K_{ie}	Inhibition constant	10 $[gL^{-1}]$
O_{sat}	Oxygen saturation concentration	2.53 x 10^{-4}[molL^{-1}]
$v_{g,max}$	max uptake flux of glucose	0.02 [mol g^{-1} h^{-1}]
$v_{o,max}$	max uptake flux of oxygen	0.008 [mol g^{-1} h^{-1}]

Study case 2: Parameter estimation

The genome-scale *S. cerevisiae* model Yeast version 8.30 (LU et al., 2019) was also applied. The uptake rates of the substrates were fixed to solve the FBA problem, the uptake rate of oxygen was set to zero, and the objective function was the maximization of biomass yield. For the anaerobic fermentation of glucose and xylose by *S. cerevisiae*, the dFBA model was formulated as follows:

$$\begin{aligned} \frac{dX}{dt} &= \mu X & \frac{dG}{dt} &= v_g X \\ \frac{dE}{dt} &= v_e X & \frac{dZ}{dt} &= v_z X \\ v_e,\ \mu &= \Xi(v_g, v_z) & v_g &= -v_{g,max}\frac{G}{K_g + G} \\ X, G, Z, E &\geq 0 & v_z &= -v_{z,max}\frac{Z}{K_z + Z}\frac{1}{1 + (G/K_{ig})} \end{aligned} \tag{3.7}$$

where μ, v_g, v_z, and v_e are the growth rate, and the exchange fluxes of glucose, xylose, and ethanol, respectively. X, G, Z, and E represent the biomass, glucose, xylose, and ethanol concentrations, respectively. $v_{g,max}$ and $v_{z,max}$ are the maximum uptake rate for glucose and xylose respectively. K_g and K_z are the saturation constants, and K_{ie} is the glucose inhibition constant. The exchange flux of ethanol v_e and cellular growth μ are computed by the mapping Ξ, which can be solved either by the optimization FBA problem (Equation 3.1) or by the surrogate approximation (Equation (3.5)).

The methodology to generate the surrogate FBA model was performed as described in Study case 1 (OLIVEIRA et al., 2021a). First, a series of optimization problems (FBA) were

solved covering the whole flux input domain (v_g and v_z). After that, a polynomial model for each output (μ and v_e) was fitted to the data by Partial Least Square (PLS) to avoid over-fitting. A parameter estimation problem for estimating the five parameters in Equation 3.7 was implemented ($v_{g,max}$, $v_{z,max}$, K_g, K_z and K_{ie}). The measurements data of the extracellular metabolites were generated by *in silico* experiments using the nominal parameter values presented in Table 11. The parameter estimation was solved as a nonlinear constrained least-squares problem as follow:

$$min_\theta \sum_j (y_j^c(\theta) - y_j^m(\theta))^2 \qquad \text{subject to: Equation 3.7} \tag{3.8}$$

where θ is the vector of the parameter to be estimated, and y is the vector of extracellular concentrations. Indexes c and m indicate calculated and measured quantities respectively. Three different methods were applied to solve the problem in Equation 3.8:

1. **dFBA + *lsqnonlin*:** Solved as a bi-level optimization problem. The outer parameter estimation problem was solved by *lsqnonlin* routine in MATLAB with the levenberg-marquardt method. The ODE system was solved with ODE15s with the embedded LP (FBA) solved in GUROBI. No gradient information was supplied.
2. **dFBA surrogate + *lsqnonlin*:** Solved as a single-level optimization problem by *lsqnonlin* routine in MATLAB with the levenberg-marquardt method. The ODE system was solved with ODE15s with the embedded LP (FBA) replaced by the surrogate model. No gradient information was supplied.
3. **dFBA surrogate + IPOPT:** Solved as a single-level optimization problem in Julia language with the interior point NLP solver IPOPT. The ODE system was solved by orthogonal collocation with the embedded LP (FBA) replaced by the surrogate model. Automatic differentiation package was used to compute the gradient.

3.3 Results

Study case 1: Economic MPC

Model structure identification

As discussed previously, the goal of the surrogate models is to approximate the rigorous solution of the mapping Ξ in Equation (3.6). Here, we use the polynomial functions

of Equation (3.4) to approximate the growth rate μ and exchange flux of ethanol v_e profiles, respectively.

In order to define the structure of the polynomial approximations, we solved the FBA problem in Equation (3.1) for different values of the uptake rates of glucose v_g and oxygen v_o. These values are defined based on an equidistant 50 by 50 grid, in which the independent variables range from 0 and to the maximum values of the constants in Table 5, i.e. $v_{g,max} = 0.02$ [mol g^{-1} h^{-1}] and $v_{o,max} = 0.008$ [mol g^{-1} h^{-1}]. The profiles of the FBA solutions are shown in Figure 14.

As expected, the increase of glucose and oxygen uptake rates have, in general, a positive impact on the growth rate. For ethanol production, the highest rates are in the anaerobic regime, i.e. no oxygen uptake. First, a unique polynomial for the whole domain was fitted to each surface (Appendix: Figure 34), however, the Root Mean Square Error (RMSE) and the Root Mean Squared Error of Prediction (RMSEP) presented unacceptable values (Appendix: Figure 35), mainly for the growth rate model. In spite of the fact a reasonable result in the controller was obtained with a single polynomial (Figure 36 and Figure 37), we decided to divide the surfaces into zones and use piecewise polynomials, with a smoothed transition.

Based on the response surface, three operation zones were formulated (Figure 14). These operation zones are analogous to the concept of phenotypic phase plan presented by Duarte et al. (2021). However, in this work, the zones were chosen based on their curvature.The operation zones were divided by visual inspection for that case study; however, a more systematic approach must be investigated in future works. The application of state of the art machine learning techniques like regression and classification trees would be interesting. However, a unique model, flexible enough could have been chosen instead.

The first operation zone (blue) is when glucose and oxygen are at higher rates. This zone represents the maximum biomass formation because of the high oxygen availability. The second zone (red) occurs when the glucose uptake rate is low compared to the oxygen uptake rate. Finally, the third zone (yellow) encompasses the anaerobic regime when the cells have to increase the synthesis of ethanol that is necessary to maintain internal redox balance. Zones red and blue are growth zones, and zone yellow is a production zone. The objective of the controller was balancing between this two operation modes to ensure a high productivity.

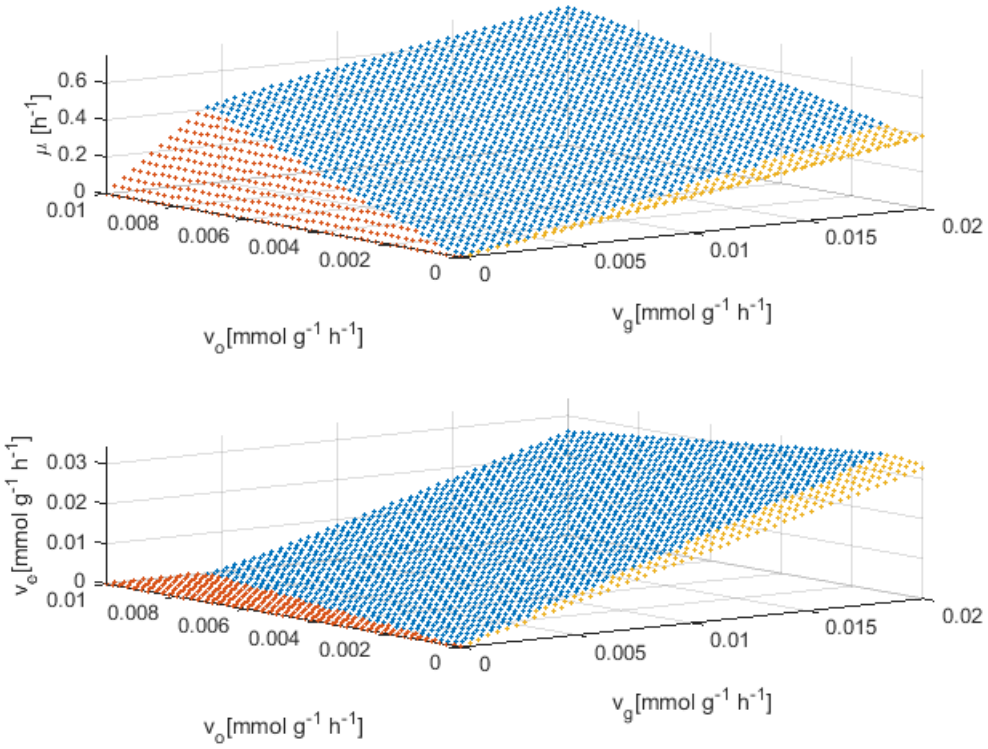


Figure 14 – Profile of the FBA solutions for different values of the uptake rates of glucose v_g and of oxygen v_o. We solved the FBA problem for every value of the independent variables in an equidistant 50 by 50 grid.

With the FBA solution profiles available and the operation zones defined, the next step for obtaining the surrogate models was to identify the polynomial parameters. We carried out a PLS analysis for comparing the different model options. The purpose of applying PLS regression is to evaluate the hyper parameter corresponding to the number of PLS components retained. This hyper parameter is able to express the balance between over-fit and under-fit.

For the model structure comparison, we consider the Root Mean Square Error (RMSE) to evaluate how well a given model fits the training data set, and the Root Mean Squared Error of Prediction (RMSEP) of a 10-fold cross-validation analysis to indicate the lack of prediction accuracy. Both indexes are normalized by the mean of the maximum and minimum value of each model, i.e. 0.374 for μ and 0.017 for v_e.

We performed the analysis using a relatively large polynomial with 20 parameters and then use PLS analysis to determine the number of PLS components to retain. The

results of the average of RMSE and RMSEP for the three zones are shown on percentage in Figure 15.

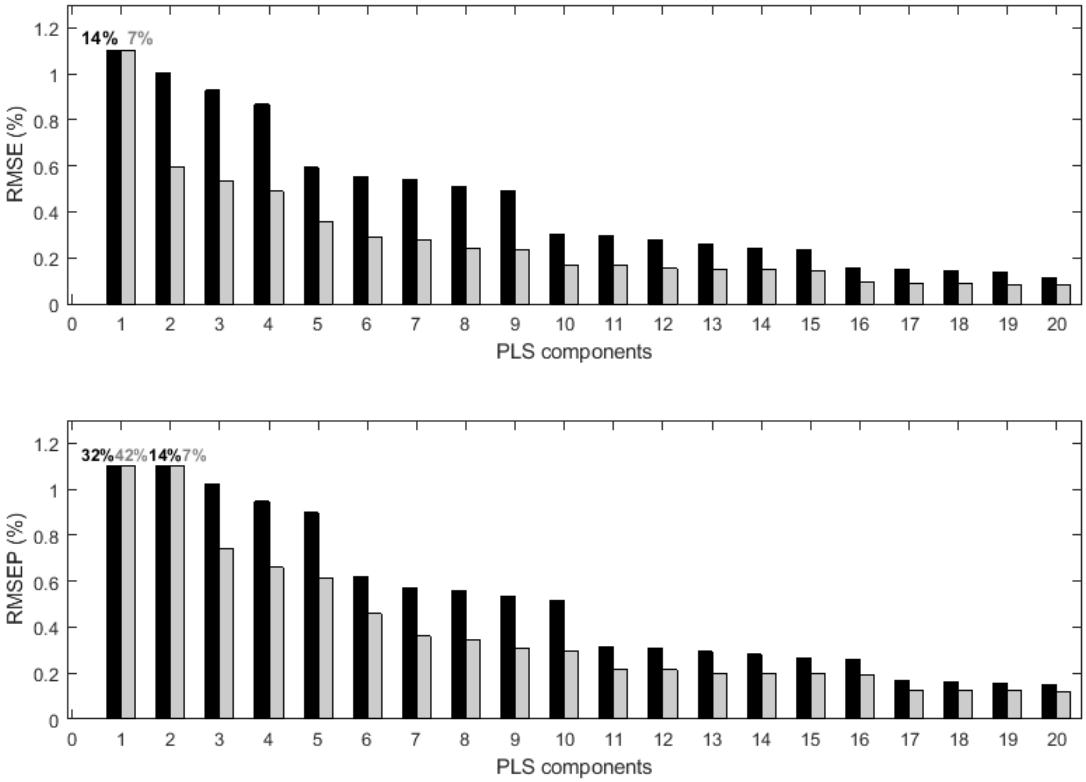


Figure 15 – Relative RMSE and RMSEP for growth rate μ model (black), and the exchange ethanol rate v_e (grey) as a function of the number of PLS components.

The results indicate that the models with only two PLS components have a poor prediction capacity. Both models for μ and v_e have a similar tendency in RMSE and RMSEP when the number of PLS components increase. However, as a consequence of a more flat response surface (shown in Figure 14), the model for v_e presents a better fitting and prediction.

Defining a threshold to decide the number of PLS components can be challenging(GELADI; KOWALSKI, 1986). Normally, heuristics are used to establish an acceptable threshold for RMSEP or a minimum variation after the addition of a component. Here, using more than 11 PLS components marginally improves the model prediction capacity and fitting for both the growth rate μ and the exchange ethanol v_e. Since an RMSEP of 0.4 % is acceptable for our proposes, the response surfaces are approximated by models obtained by using 11 PLS components.

Model validation

After determining the surrogate model, we need to assess how it influences the complete reactor model accuracy. The comparison of the solution of Equation (3.6) using the mapping Ξ as the FBA problem in Equation (3.1) and the surrogate approximation in Equation (3.4) was performed. The parameters of Table 5 were used in this simulation.

Since the surrogate model consists of three polynomial models (one for each zone in Figure 13) connected by a smoothed heaviside step function, preliminary simulations were used for tuning the heaviside parameter (k_l). This parameter controls how sharp is the transition between models. The adjusted heaviside parameter was $k_l = 1.0x10^3$, which led to a smooth transitions between the operation zones.

The validation simulation starts with DO = 0 % (yellow zone in Figure 13), then, after 6h the DO is changed to 50 % (blue zone in Figure 13), finally, as glucose is being consumed, the process reaches the third zone (red in Figure 13). The results are shown in Figure 16. The simulation for standard dFBA (dashed) and surrogate dFBA (dotted) present overlapping profiles.

The simulations were performed in JULIA language (BEZANSON et al., 2017) using an explicit Runge-Kutta method from JULIA's DifferentialEquations package (RACKAUCKAS; NIE, 2017). Both standard dFBA and surrogate dFBA were solved by the DA method. However, the standard dFBA was solved in about 120 s, while the surrogate dFBA was solved in 4 s in an Intel core i7 @ 1.8GHz . Since the LP/FBA problem needs to be solved at each integration step, it makes the DA method for the standard dFBA time-consuming. For comparison, the solution of the same problem for the standard dFBA with the SOA method was solved in about 60 s. For simulations of the dFBA model alone, it is evident that the use of the state-of-the-art methods would be preferential, not because of the time of simulation, but because of the time needed to train the surrogate model. However, if the aim is to insert the dFBA model on an architecture such as optimal control or parameter estimation problems then, in our opinion, the surrogate dFBA would culminate in a single-level optimization problem with a fast computation of gradient and hessian information.

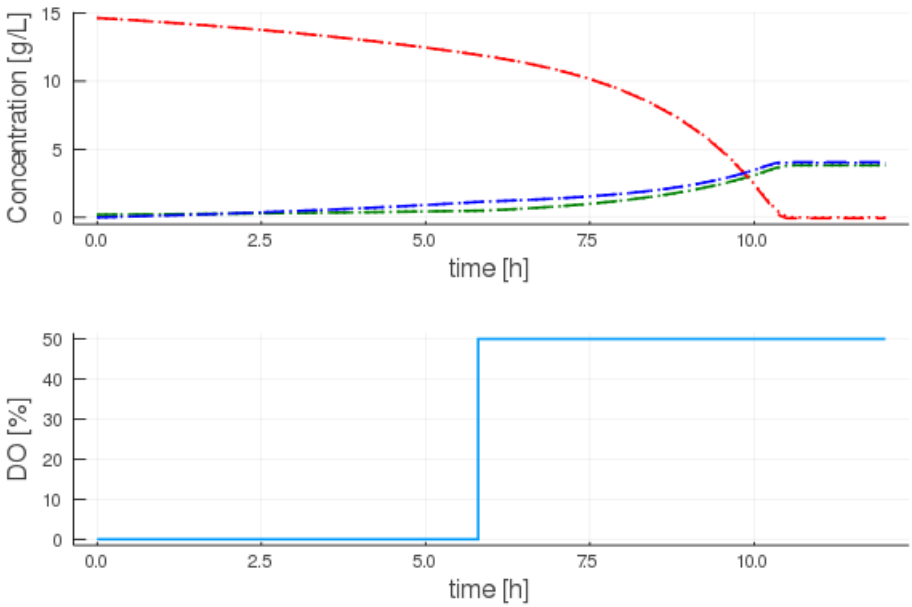


Figure 16 – Comparison between simulated profiles for standard DFBA (dashed) and surrogate DFBA (dotted). Glucose (red), biomass (blue) and ethanol (green).

Controller formulation using the surrogate model for dFBA

The EMPC controller is formulated to maximize ethanol mass at the end of the control horizon. The manipulated variables are F and DO. In order to simulate the case study problem, we use the model described in Equation (3.6) both as the "plant" and as the controller model. They differ on how the mapping Ξ is solved. For the "plant", we use the FBA formulation of Equation (3.1), whereas our method with the surrogate model approximation is used in the controller (Equation (3.5)). The EMPC formulation, then, becomes:

$$
\begin{aligned}
&\max_{u(t)=[F(t),DO(t)]} \quad \phi := V(t_0+T_p)E(t_0+T_p) \\
&\text{s.t. } \textit{DFBA model:} \\
&\dot{y}(t) = f(y(t),u(t)), \quad \text{for} \quad t \in (t_0, t_0+T_p] \\
&y(0) = y_0 \\
&\textit{inequality constraints:} \\
&V(t) \leq 1.2 \\
&F(t) \geq 0 \\
&0 \leq DO(t) \leq 50
\end{aligned}
\tag{3.9}
$$

where $y(t) = [V(t), X(t), G(t), E(t)]^T$, and $u(t) = [F(t), DO(t)]^T$. The dFBA model is represented by an ordinary differential equation (ODE), which results from rearranging Equation 3.6. The inequality constraints incorporate technical restrictions of the system. We assume that the final simulation time is $t_{end} = 13[h]$, t_0 is the current time, the prediction horizon is $T_p = 50[min]$, and the controlling sampling time is $h = 10[min]$. In this control formulation, we also constrained the number of control moves by including a control horizon of $T_c = 30[min]$. Full state feedback is assumed in the simulations

The controller is implemented in JULIA language (BEZANSON et al., 2017) and IPOPT (WÄCHTER; BIEGLER, 2006) is used for solving the optimization problems. The “plant” bioreactor model is integrated in time using an explicit Runge-Kutta method from JULIA’s DifferentialEquations package (RACKAUCKAS; NIE, 2017) (DA method). On the other hand, the differential equations of the controller model are discretized using Radau collocation on finite elements with 3 collocation points in each interval, and integrated by the NLP solver (DOA method).

EMPC implementation

The economic NMPC implementation is evaluated in two steps: First, we evaluate the controller performance in the absence of modeling errors, i.e. the dFBA with the surrogate model is used as the true model of the bioreactor. Second, we add plant-model mismatch to the analysis by considering errors in the surrogate model identification process.In this case, the “plant” bioreactor model is set as the standard dFBA, in which the linear optimization is performed in the genome-scale network model *null space*. On the other hand, the parameters of the surrogate model are estimated from the FBA simulations with the genome-scale network model, leading to the model discrepancy.

No plant-model mismatch

Figure 17 shows the results for the scenario without plant-model mismatch. The results are presented in the following order: the top left plot shows the evolution of the volume inside the fed-batch reactor as well as the maximum volume constraint; the top right plot shows the concentrations of glucose (red), biomass (blue) and ethanol (green). The

latter is the product of interest of the reaction being carried out. The bottom plots show profiles of the manipulated variable profiles, feed flowrate (left) and dissolved oxygen (right).

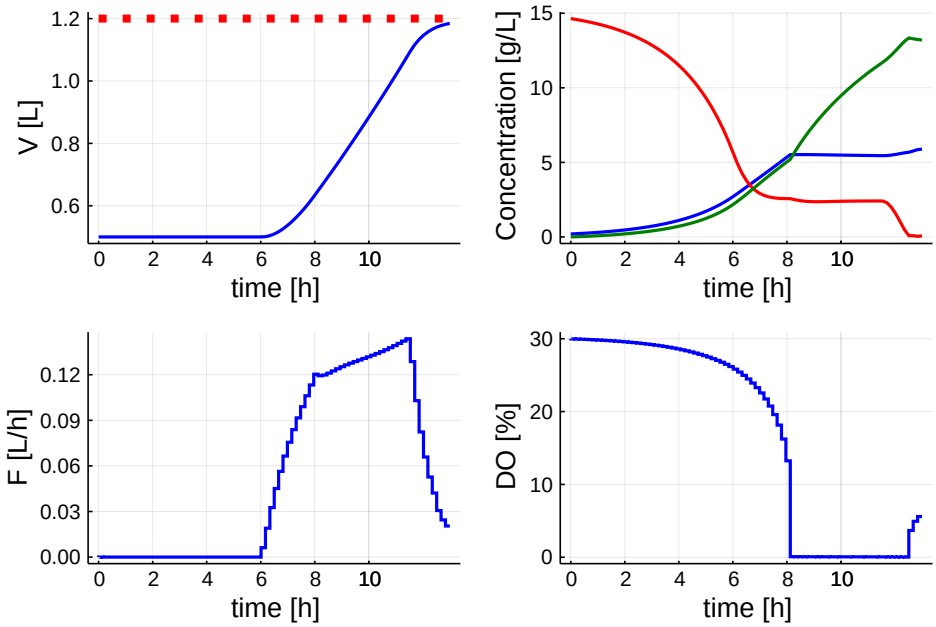


Figure 17 – Profiles obtained using the surrogate model both in the controller and as the "plant" bioreactor model. The concentration profiles are glucose (red), biomass (blue) and ethanol (green). In the volume plot, the maximum volume constraint is shown in red.

The final obtained ethanol mass is 15.64 [g] with a final biomass mass of 6.97 [g]. The glucose feeding starts at 6 [h] of batch time and the aerobic-anaerobic switch happens after 8 [h]. Additionally, the controller keeps the glucose concentration in an optimal value (~ 2.4 [$g\ L^{-1}$]) during the ethanol production phase, which is important for the reaction yield. If the glucose concentration is high, it inhibits the microorganisms activity. On the other hand, if it is low, the ethanol production decreases.

The profiles are similar to the results reported in Chang et al. (2016), where the case study was originally introduced. The direct comparison cannot be performed because the authors used a smaller scale network model. However, this qualitative result indicates an advantage of our method. In Chang et al. (2016), the FBA LP problem first order optimality conditions are added as constraints in the controller formulation (Equation (3.9)), which allows the simultaneous solution of both optimization problems (LP (FBA) and control). This strategy requires a reduction model step, in which the network model is simplified, or a small-scale network model is used. Since network models are getting larger and more detailed (LLOYD et al., 2018), this simplification step may discard important information about the

underlying metabolic pathways and affect the controller performance. In turn, in our method, the surrogate models capture the entire information provided by the detailed genome-scale network model, which can be explored by the controller.

Implementation issues

In a preliminary study, before obtaining the results of Figure 17, the controller promoted abruptly changes in the manipulated variables (F and DO) during the simulations. This policy culminated in glucose concentration fluctuations, which are not desirable in real implementations. The same patterns were reported by Chang et al. (2016). The authors solved the problem by imposing constraints in the manipulated variables. However, this strategy did not have the same impact in our simulations. Then, in order to obtain smoother manipulated variable profiles, a penalty was added in the controller objective function (see Equation 3.9), which became:

$$\phi := V(t_0 + T_p)E(t_0 + T_p) + \int_{t_0}^{t_0+T_p} \dot{\mathbf{u}}(t)^T \mathbf{R}\dot{\mathbf{u}}(t)\, dt \tag{3.10}$$

where $\dot{\mathbf{u}}(t) = [\dot{F}(t), \dot{DO}(t)]^T$ are the input changes and $\mathbf{R}$ is a weighting matrix with appropriate dimensions. Here it was set as 0.1.

Another challenge during the control implementation was related to the NLP convergence. Initially, the control problem did not converged in approximately 10% of the times. The problem was solved by relaxing some equality constraints using slack variables and adding one more penalty in the objective function. The new control formulation, then, became:

$$\max_{\mathbf{u}(t)=[F(t),DO(t)]^T} \quad \phi := V(t_0+T_p)E(t_0+T_p) + \int_{t_0}^{t_0+T_p} \dot{\mathbf{u}}(t)^T \mathbf{R}\, \dot{\mathbf{u}}(t)\, dt + \int_{t_0}^{t_0+T_p} \mathbf{s}(t)^T \mathbf{Q}\, \mathbf{s}(t)\, dt$$

s.t. *DFBA model:*

$$\dot{\mathbf{y}}(t) = \mathbf{f}(\mathbf{y}(t), \mathbf{u}(t)), \quad \text{for} \quad t \in (t_0, t_0+T_p]$$

$$y(0) = y_0$$

$$\mathbf{s}(t) = [s(t)^F, s(t)^{DO}]^T$$

$$\mathbf{y}(t) = [V(t), X(t), G(t), E(t)]^T$$

inequality constraints:

$$V(t) \leq 1.2$$

$$F(t) \geq 0$$

$$0 \leq DO(t) \leq 50$$

$$|\dot{F}(t)| \leq 0.02 + s(t)^F$$

$$|\dot{DO}(t)| \leq 0.5 + s(t)^{DO} \tag{3.11}$$

Parameter mismatch

Although the surrogate modeling approach has advantages, a poor approximation may affect the control performance because of plant-model mismatch. In order to illustrate this scenario, we use a less accurate version of the surrogate model inside the controller. Instead of using 11 PLS components, as in the previous section, we use only 2. As a consequence, the average prediction error of the surrogate models increases to approximately 14% (see Figure 15), leading to a significant degree of plant-model mismatch.

For testing the controller with the less accurate surrogate model, we run the controller both in open loop (without state feedback) and in closed loop (considering full state feedback). Feedback plays an important role in the control performance providing robustness to modeling uncertainty (ASTROM; MURRAY, 2008). If the model response deviates significantly from the true system response, which indicates plant model-mismatch, feedback supplies a corrective action to the model. In turn, if no feedback is used, the control performance heavily depends on the model accuracy. If there is any degree of plant-model mismatch, it is very likely that the control results are poor.

Figure 18 illustrates how we can use feedback to deal with the plant-model mismatch created by less accurate surrogate model approximations. The figure setting is the same as in Figure 17. We plot the responses of the open loop (dotted line) and closed loop (full line) together.

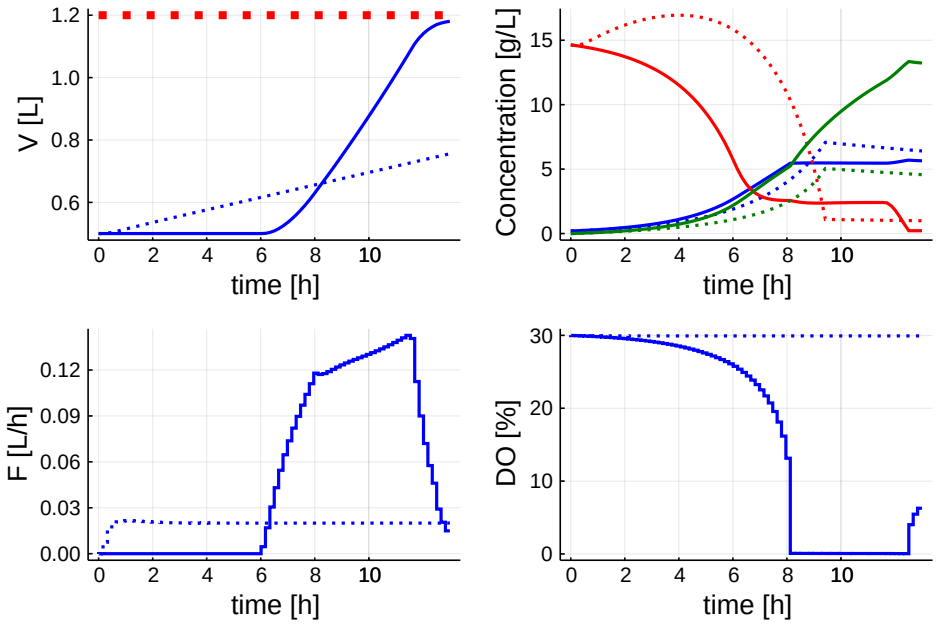


Figure 18 – Comparison between open (·) and closed loop (–) operation. The concentration profiles are glucose (red), biomass (blue) and ethanol (green). Here, the surrogate model is obtained by using the first two PLS components, while the "true" system is based on genome-scale network null space.

The results show that, in this case, the controller feedback mitigates the effects of the plant-model mismatch. The final ethanol concentration obtained was 15.60 [g] with a final biomass mass of 6.66 [g]. Both are are very similar to the results shown in the previous sections, where the controller model it is a surrogate model. The final ethanol concentration deviates only 0.25% of the solution with the perfect control, whereas the final biomass concentration deviates 4.6% .

In contrast, the open-loop had a poor performance. The open-loop formulation was unable to follow the behavior of the real model. The controller overestimates the glucose uptake rate and, as a result, glucose started to accumulate in the bioreactor, inhibiting the ethanol production.

Remark. *For completeness, we performed two extra simulations, in which the controller model is different from the plant. We used two distinct genome-scale models (GSM) as the plant models (the GSM Saccharomyces cerevisiae models iND750 and iMM904), whereas*

the controller model was trained using Yeast 8.3. In both cases the plant-model mismatch impact on the controller performance was small. For the sake of brevity, the results are shown in the Appendices (Figure 38).

Noisy measurements

A test to verify the performance of the controller in the presence of noise in measurements was carried out. A zero-mean white noise of 10% was added to the biomass, glucose, and ethanol measurements, and 1% for volume measurements. The goal of this test is to mimic an imperfect state feedback. The surrogate models using 11 PLS components were set to the controller and the standard dFBA for the "plant" bioreactor model. 100 simulations were performed to evaluate the performance distribution of the closed-loop controller (Figure 19). It is important to say that measurements were directly used in the controller, without any filter. The profiles do not considerably differ from the original trajectories (–). Furthermore, In all simulations, the controller guaranteed that the operational constraints were not violated. The standard deviation in the final ethanol concentration from the perfect control case was 1.18%, and the worst-case deviation was 3.33 %.

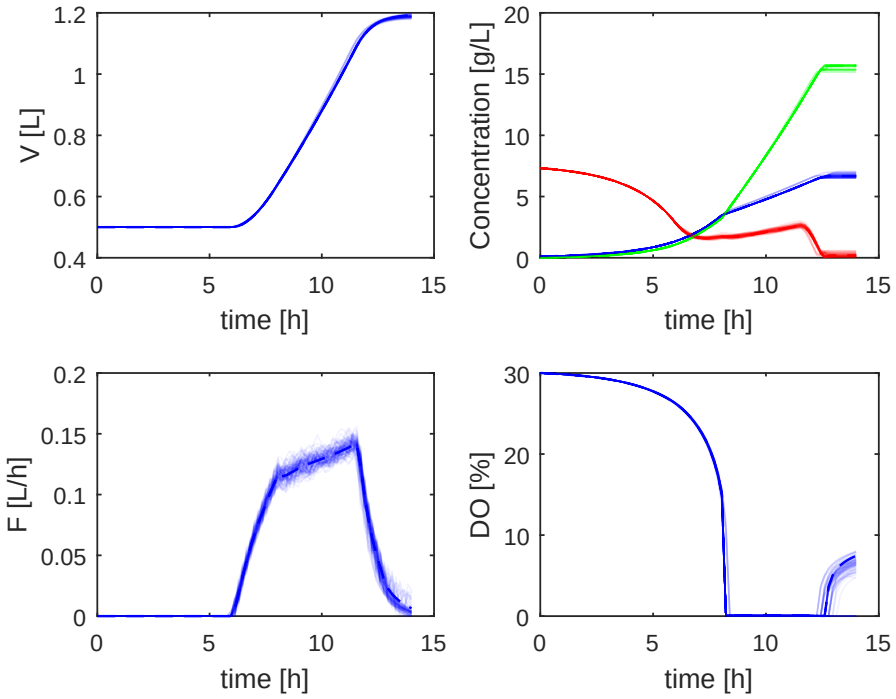


Figure 19 – Comparison between 100 simulations of closed loop operation with noisy measurements. The concentration profiles are glucose (red), biomass (blue) and ethanol (green). Here, a zero mean white noise of 10% was added to the biomass, glucose and ethanol measurements, and 1% for volume measurements. (–) reference profiles without noisy measurements.

Both Figure 18 and Figure 19 show that feedback can alleviate the plant-model mismatch if it is present. On the other hand, for a proper feedback, we need sensors that provide accurate measurements, with proper levels of measurement noise. It is crucial to properly balance the trade-off between the costs of sensors with obtaining a more detailed model and the benefits that the controller can bring. Also, it is important to consider that as a bottom-up model the process of training the surrogate model from a detailed genome scale description is much more simple than the estimation process for top-down models, that could be obtained from a large number of experiments.

Study case 2: Parameter estimation

First, the surrogate FBA model was trained using FBA simulations performed in COBRA Toolbox for MATLAB. The FBA was solved for every value of the v_g and v_z uptake fluxes in an equidistant 40 by 40 grid (Figure 20). Both response surfaces for μ and v_e are flat, which means that the amount of ethanol and biomass being produced are linearly proportional to the uptake of each substrate. Different from the non-linear response surface

in Study case 1 (OLIVEIRA et al., 2021a) where a piecewise polynomial surrogate model was needed, here a single polynomial could fit the data. The relative Root Mean Square Error (RMSE) was $6.67e-8$ % and $1.41e-4$ % for μ and v_e respectively. While the relative Root Mean Square Error of Prediction (RMSEP) was $2.99e-10$ % and $1.91e-6$ % for μ and v_e respectively. The relative RMSE and RMSEP were computed by dividing the fluxes by the maximum value of the correspondent uptake flux.

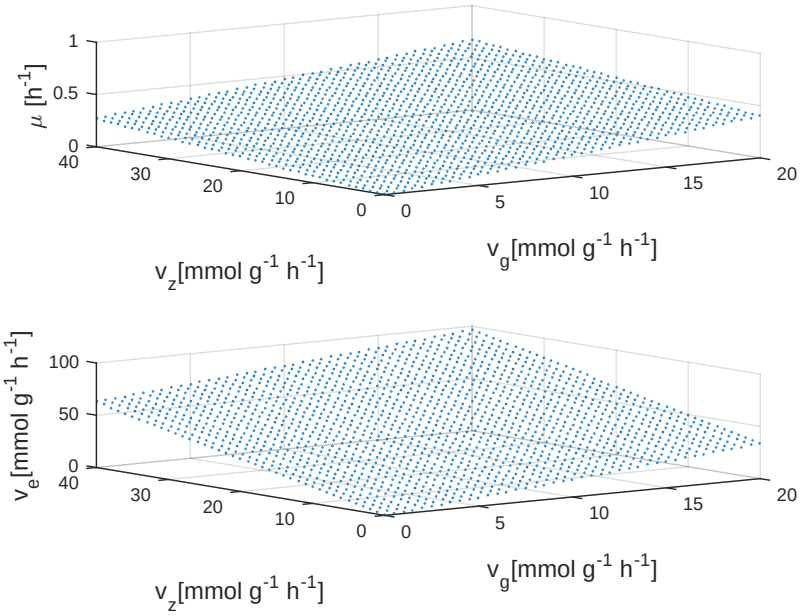


Figure 20 – Profiles of the FBA (Equation 3.1) solutions for different values of the uptake rates of glucose v_g and xylose v_z. We solved the FBA problem for every value of the independent variables in an equidistant 40 by 40 grid.

After the FBA surrogate model was trained, the parameter estimation problem was solved using the three different methods described in the methodology section. Ten different initial guesses were supplied to solve the problem by each method and the solution with the lower objective function (OF) was selected. Figure 21 compares the predicted concentrations with the best-fitted set of parameters for each case. Visually, the methods that applied the surrogate FBA fit the *in silico* data adequately, on the other hand, the method that uses the nested LP to solve the dFBA model was unable to fit the data. In fact, all the attempts to solve the estimation problem using the nested LP resulted in a set of parameters close to the initial guess. Because of the embedded optimization problem, the *lsqnonlin* solver was not able to compute efficiently the gradient of the problem. Furthermore, the attempts of using derivative-free methods like simplex (i.e. *fminsearch*) have failed as well.

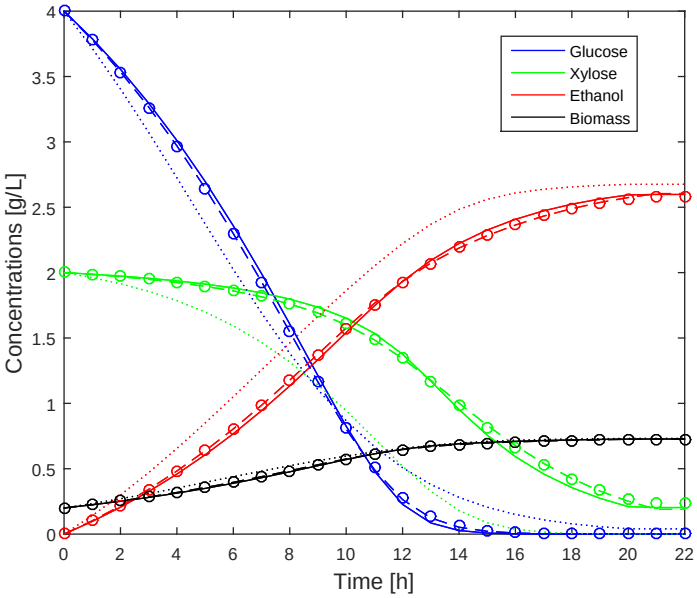


Figure 21 – Simulated profiles for dFBA surrogate in Julia (solid), dFBA surrogate in MATLAB (dashed) and dFBA DA in MATLAB (dotted). *in silico* data points are presented as circles.

The performance of each method for solving the problem is presented in Table 11. The dFBA with the embedded optimization had a CPU time about 60 times higher than the methods that used the surrogate FBA. Despite the larger CPU time, the solver performed only 5 iterations and 36 function evaluations. The need of solving the nested LP at each step of the ODE solver makes this method computationally expensive and ineffective, as it can be seen by the poor fit as well (Figure 21). Comparing the methods that applied the surrogate approximation, the utilization of automatic differentiation can improve performance. However, the utilization of the surrogate FBA is enough to guarantee a good fitting. These results illustrate the advantage of the surrogate approximation FBA to solve parameter estimation problems. The time and effort to train the surrogate model must be taken into account in that analysis, but for a small number of input fluxes the task can be easily done. Moreover, Table 11 also presents the set of parameters estimated in each case, as well as the set of parameters used to generate the measurements. The set of estimated parameters was different from the one used to simulate the measurements data even when a good fit was achieved. In fact, the set of parameters in dFBA models are typically dependent and cannot be uniquely identified (LEPPäVUORI et al., 2011). In order to make a complete analysis, parameter uncertainty must be taken into consideration. This is a practical identifiability issue that should be discussed with more depth. A possibility is to apply a recent methodology developed in our group that used sparse Principal Component Analysis to access the identifiability of metabolic fluxes on carbon labeling experiments (OLIVEIRA et al., 2021b).

Table 6 – Comparison of the Computational Performance of each method and the parameter values used in model simulation to yield measurements.

	Model simulation	dFBA surrogate IPOPT	dFBA surrogate *lsqnonlin*	dFBA *lsqnonlin*
CPU	-	0.05 s	26.47 s	27.82 min
iterations	-	29	20	5
function evaluations	-	50	126	36
Objective function	-	0.049	0.009	7.17
v_g^{max}	7.30	6.44	7.13	30.01
K_g	1.03	0.64	0.94	12.02
v_z^{max}	32.00	26.07	4.69	7.99
K_z	14.85	10.48	1.60	0.80
K_{ie}	0.50	0.39	0.81	1.00

3.3.1 Conclusions

We presented a new NMPC application to bio-reactors. By combining mass balances with a surrogate model for the genome-scale network, our approach provides significant reduction in online computations while still accurately describing the microorganism of interest.

The core idea is to replace the inner LP problem associated with the solution of the flux balances analysis (FBA) of the genome-scale network model by a polynomial. Therefore, the computational issues associated with solving a bi-level optimization problem, which comes from placing the LP inside the optimal control problem, are avoided. Moreover, apart from the approximation errors due to the use of a surrogate model, our approach does not require any further simplification of metabolic networks. In this work, we suggested an approximation identification procedure based on Partial Least Squares Regression to assure that the polynomials are an accurate representation of the underlying metabolic network.

Our approach is an alternative to the controllers that require a reformulation of the FBA problem (CHANG et al., 2016). In these controllers, instead of directly solving the FBA, the problem first-order optimality conditions are used as constraints to the optimal control problem. The results show that the NMPC with the surrogate models has an equivalent performance, in terms of manipulated variables profiles and economic results, to the controller with reformulated FBA reported by Chang et al. (2016). Moreover, the controller used in Chang et al. (2016) relies on small scale genome-scale metabolic models whereas our

approach relies on polynomial approximations of the complete metabolic network, which are more representative of the underlying system.

The strategy was also applied to a parameter estimation study case. The estimation of the model parameters poses a challenge due to the bi-level optimization architecture and the non-smooth behavior of the dFBA model. Here, the replacement of the embedded optimization problem by a surrogate model was evaluated. The results demonstrated that the surrogate model can be easily trained from FBA simulations and improve the performance of the estimation problem. In the future, the methodology should be applied to parameter estimation and uncertainty quantification problems using experimental data.

Finally, we would like to summarize some limitations of the current work, which are interesting for future research. The FBA approach can give rise to a non-unique set of solutions (ORTH et al., 2010), which can be more common in genome-scale models that are not completely validated. Another aspect that should be addressed in the future is the case when not all state measurements are available and some of them must be estimated. An approach similar to Jabarivelisdeh et al. (2020) could be used, where a moving horizon estimation and resource balance analysis algorithms were combined with the MPC in order to estimate the states and account for parameter uncertainty. Additionally, for more complex examples, the polynomial models could not be a good fit, then other surrogate models families should be explored.

4 Nonlinear Programming Reformulation of Dynamic Flux Balance Analysis Models

The description of the metabolism dynamic behavior is relevant for a range of research fields such as systems biology, synthetic biology, metabolic engineering, and bioprocess engineering (NIELSEN, 2017). Dynamic models have been used for a variety of applications such as predicting gene deletion effects on a bioproduct productivity (HJERSTED et al., 2007), optimal control of bioprocesses (CHANG et al., 2016), and understating the effect of diseases on human cell metabolism (GUEBILA; THIELE, 2021). A widely applied dynamic model of metabolism is the so-called Dynamic Flux Balance Analysis (dFBA) (MAHADEVAN et al., 2002). The main hypothesis in dFBA models is that the intracellular dynamics are faster than the extracellular dynamics, in consequence the intracellular metabolites could be described by a steady-state model. In order to implement that hypothesis, a metabolic network is used. The networks are built using omics information, mapping transformations between substrates to bioproducts mediated by enzymes (THIELE; PALSSON, 2010). The metabolic networks are expressed in the form of a matrix (stoichiometric matrix), in which each column corresponds to an enzymatic reaction and each row to a metabolite. Because typically those matrices have more fluxes than metabolites, an undetermined system of equations is to be solved. An optimization problem is formulated for that purpose using a biological meaningful objective function. This process is called Flux Balance Analysis (FBA) (ORTH et al., 2010). The dynamic version of FBA is implemented by applying mass balances equations to describe the behavior of the extracellular metabolites, in conjunction with kinetic expressions aimed to describe the uptake of substrates. The solution of dFBA models present some challenges that are summarized below:

- Stiff behavior: Normally dFBA models present stiff behavior, especially in situations as diauxic growth (MAHADEVAN et al., 2002; HJERSTED et al., 2007).
- Non-unique solutions: in most cases, the solution of the optimization problem (FBA) is not unique. In consequence, situations where the fluxes ”jump” between different optimal solutions can arise.
- Feasibility: When the optimization solver cannot find a feasible solution for the FBA problem, the interaction with an ODE solver can be tricky.
- Scalability: The size of the metabolic networks applied in FBA can be large, therefore methods should scale adequately for those situations.

- Nonlinearity: The presence of nonlinear constraints or objective functions can substantially increase the computational demand.
- Differentiability: When the dFBA models are used inside an optimal control or parameter estimation architecture, the first and second derivatives of the model must be computed.

The simplest method for solving a dFBA model is the Static optimization approach (SOA) (MAHADEVAN et al., 2002). In the SOA the ODE system is discretized using an explicit Euler method and the FBA is solved at each time step. For stiff problems, the Euler method would need a small step, then the FBA needs to be solved a large number of times. Another usual approach is the direct approach (DA) (ZHUANG et al., 2011), where an adaptive size solver for the ODE system is used and the number of optimizations decreases in comparison with the SOA method. Both SOA and DA are easy to implement but the solution of the optimization problem can make the simulations fail.

Modifications on the DA have been proposed in order to address this issue, like event-based methods. Gomez et al. (2014) developed an event-based methodology that detects changes in active-set of the FBA model (HARWOOD et al., 2016). This makes the solution faster because the FBA problem does not need to be solved at each time step. They also applied a lexicographic optimization methodology to deal with the non-unique solutions of FBA, wherein the first optimization problem solves a feasibility problem. Other event-based methods that compute an optimal basis for the FBA feasible space have also been proposed (PLOCH et al., 2020; BRUNNER; CHIA, 2020). The drawback of these formulations is the need of continuous monitoring the active-set of the optimization problem that can increase with the size of the metabolic network. Also, the model is non-differentiable at the points of change of the active-set.

Another class of approaches propose the reformulation of the dFBA using the Karush Kuhn Tucker (KKT) conditions of the FBA problem. Ploch et al. (2020) reformulated the dFBA model as a nonsmooth DAE system and applied a homotopy continuation to avoid the fluxes from jumping between optimal solutions. Scott et al. (2018) uused an interior point reformulation of FBA. The solvers failed to solve the generated DAE system and they needed to apply DAE index reduction to obtain an implicit ODE system. Dynamic optimization problems using genome-scale networks were solved and the model derivative information was computed solving the sensitivity equations. Surrogate models of dFBA have

also been proposed, they focused mainly on Model Predictive Control (MPC) applications, for introducing robustness using a polynomial chaos expansion (KUMAR; BUDMAN, 2017) or to allow the resolution of large genome-scale models in a short time and high convergence rate (OLIVEIRA et al., 2021a).

Finally, there is a class of approaches that reformulate the problem as a nonlinear programming (NLP) problem. These approaches discretize the ODE system using a technique such as orthogonal collocation on finite elements, which are included as constraints in the optimization problem. The simplest version is the Dynamic Optimization approach (DOA) (MAHADEVAN et al., 2002) that uses some instantaneous objective function such as the maximization of biomass formation. Raghunathan et al. (2003) formulate a parameter estimation problem for dFBA problems using the KKT conditions for the FBA problem. Chang et al. (2016) did the same for MPC formulation. The great advantage of these NLP formulations is that they can easily incorporate path constraints, nonlinear constraints, nonlinear objective functions and they can be easily integrated into parameter estimation and optimal control problems as a single-level problem. Also, the FBA optimization does not need to be solved a large number of times during the simulation. The main drawback of this strategy is the need to solve large NLP that can be difficult to initialize and can suffer from convergence problems (MAHADEVAN et al., 2002; CHANG et al., 2016).

Here, we present an NLP reformulation for the dFBA models that is fast, precise, and can be applied for large genome-scale networks. The methodology consists in incorporating a KKT reformulation of a parsimonious FBA problem, applying a collocation technique to the ODE system. The method was implemented using the free open source and high-performance language Julia (BEZANSON et al., 2017). Furthermore, the utilization of automatic differentiation code packages can allow for fast computation of first and second-order derivatives. This represents a considerable advance in comparison to the simulation-based approaches that need to compute the derivatives by finite differences, sensitivities equations, or applying derivative-free methods. Also, the application of a large-scale NLP solver (e.g. IPOPT) can considerably reduce the computational time and handle large optimization problems by exploring the sparsity of the NLP systems (SHIN et al., 2019).

4.1 Methodology

ODE discretization using orthogonal collocation on finite elements

The Ordinary Differential Equation (ODE) system of the dFBA model represents the external metabolites mass balances. Efficient ODE solvers are available for stiff problems like dFBA models, and many of them have been applied until now (GOMEZ et al., 2014; SCOTT et al., 2018). However, the interaction of the optimization solver with the FBA problem can be tricky, mainly if the optimization solver does not return any solution (infeasible problem) at a given integration step. Another difficulty is that the optimization problem needs to be solved many times, at each time step of the ODE solver. For large metabolic networks or more computational demanding versions of FBA like pFBA (Equation 4.3), the sequence of optimizations can get computationally cumbersome.

Alternatively, to use an ODE solver, the ODE system can be discretized using the orthogonal collocation technique (BIEGLER, 2010) (also know as direct transcription). This technique was demonstrated to be equivalent to an implicit Runge–Kutta method with high-order accuracy and strong stability properties (BIEGLER, 2010). Here, we discretized the ODE system using orthogonal collocation on three Radau collocation points. Also, the time domain was divided into finite elements.

$$\begin{aligned} x_{k+1} &= x_k + F(x_{k+1}, v_{k+1}) \\ x(t) &\approx A + Bt + \frac{C}{2}t^2 + \frac{D}{3}t^3 \end{aligned} \tag{4.1}$$

Applying Equations 4.1 to the ODE of the state variables presented in Equation 3.7:

$$x = x_0 + hMF(x, v) \tag{4.2}$$

where M is a matrix that can be calculated based only on the position of the collocation points; h is the time step; and x_0 is the vector with the initial condition for each element that is defined as the last point of the previous element, and for the first element is defined as the initial condition of the state vector x.

Karush–Kuhn–Tucker (KKT) reformulation of FBA

In an analogous way to the ODE system, the FBA optimization problem can also be transformed into a system of algebraic equations. This can be done by applying either the duality theory (MARANAS; ZOMORRODI, 2016) or the first-order necessary optimality condition (Karush Kuhn Tucker-KKT) (RAGHUNATHAN et al., 2003). Besides the fact that the duality theory transformation avoids the complementarity constraints on the model, this formulation leads to a nonconvex NLP, for which only local optimal solutions can be guaranteed (RAGHUNATHAN et al., 2003). For this reason we decided to apply the KKT conditions here.

Typically, FBA has multiple optimal solutions, Gomez et al. (2014) applied lexicographic optimization to deal with the multiplicity of solutions, however, the number of optimization problems increases, and the order of objective functions must be supplied. Scott et al. (2018) applied an interior-point method to obtain an arbitrary unique solution in the interior of the feasible space. pFBA (Equation 4.3) has shown a good agreement with experimental data (LEWIS et al., 2010; MACHADO; HERRGåRD, 2014), therefore, here an alternative formulation for the pFBA problem presented by Ploch et al. (2020) was implemented:

$$\begin{aligned} &\max_{v} \quad c^T \boldsymbol{v} - v^T W v \\ &\text{s.t.} \\ &S\boldsymbol{v} = 0 \\ &\boldsymbol{v}_{\min} \leq \boldsymbol{v} \leq \boldsymbol{v}_{\max} \end{aligned} \tag{4.3}$$

Equation 4.3 consists in the FBA formulation with the addition of a quadratic regularization term to search for solutions with minimal overall intracellular flux. W is the regularization diagonal matrix. As observed by Ploch et al. (2020),for a small value of parameter w the solution of Equation 4.3 is also a solution of Equation 3.1. The LP FBA is then transformed into a quadratic programming (QP) problem that is also convex (PLOCH et al., 2020). The KKT conditions for the Equation 4.3 are:

$$\begin{aligned} &c - W v + S^T \lambda - \alpha^U + \alpha^L = 0 \\ &(\boldsymbol{v}_{\max} - v)\alpha^U = 0 \\ &(v - \boldsymbol{v}_{\min})\alpha^L = 0 \\ &\lambda, \alpha^U, \alpha^L \geq 0 \end{aligned} \tag{4.4}$$

where λ, α^U, and α^L are the multipliers corresponding to the stoichiometric equation, fluxes lower bound and fluxes upper bound, respectively. Equation 4.4 belongs to a class of problems called Mathematical Programs with Complementarity Constraints (MPCC) (BAUMRUCKER et al., 2008). Complementarity constraints are related constraints that at least one of them must be active. This MPCC is particularly changeling to solve because they violate the linear independence constraint qualification (LICQ), which requires the gradients of the active constraints to be linearly independent (BAUMRUCKER et al., 2008).

A series of regularization methods have been proposed to deal with the MPCC. Some methods consist in just relaxing the right-hand side of the complementarity constraints using a small value and solving a series of simulations decreasing this value (BAUMRUCKER et al., 2008). Applying an extreme-ray-based transformation (ZHAO et al., 2017), or an interior-point transformation (SCOTT et al., 2018) have also been proposed. Also, a Mixed-Integer Nonlinear Programming (MINLP) transformation can be applied (MARANAS; ZOMORRODI, 2016).

Finally, a penalization method can be applied when the complementarity constraints are inserted in the objective function as a penalty parameter. Here, we decided to apply the penalty method to deal with the MPCC because it can be solved in a single optimization problem and it has also presented good results with NLP interior-point solvers (BAUMRUCKER et al., 2008), which was applied in this work.

NLP dFBA implementation

We are now able to build an NLP formulation of the dFBA model applying the discretized ODE and the KKT conditions with the penalization method to deal with the complementarity constraints as follows:

$$
\begin{aligned}
&\max_{x,v} \quad \Phi(x,v) - \rho((\boldsymbol{v}_{\max} - v)\alpha^U + (v - \boldsymbol{v}_{\min})\alpha^L) \\
&\text{s.t.} \\
&x = x_0 + hMF(x,v) \\
&x \geq 0 \\
&v_{uptake} = g(x) \\
&lb \leq v \leq ub \\
&c - W\,v + S^T\lambda - \alpha^U + \alpha^L = 0 \\
&\lambda, \alpha^U, \alpha^L \geq 0
\end{aligned}
\tag{4.5}
$$

where ρ is the penalization parameter and $\Phi(x,v)$ is the objective function. For the case of a simple dFBA simulation, the value of $\Phi(x,v)$ is zero. However, $\Phi(x,v)$ can also be a sum of squares of residues for a parameter estimation problem, or an optimal control objective. That is one advantage of the NLP formulation, which can be easily adapted for different applications of the dFBA model. Moreover, the NLP formulation allows direct gradient and Hessian evaluations, differently from the sensitivity calculations when DAE solvers are used, which can cause convergence difficulties (BIEGLER, 2010) and requires that the dynamic model be simulated many times (SHIN et al., 2019).

On the other hand, efficient large-scale NLP solvers are required for efficient solution of the NLP formulation. Therefore, we implemented the NLP dFBA formulation on JULIA language (BEZANSON et al., 2017) using the automatic differentiation package (RACKAUCKAS; NIE, 2017) for fast computation of the gradient and Hessian. The NLP was solved using the IPOPT solver (WÄCHTER; BIEGLER, 2006) with the linear solver MA27. An overview of the NLP dFBA approach can be found in Figure 22. All calculations were performed on a laptop equipped with an Intel Core i5-1135 CPU, 2.40 GHz and 16 GB RAM.

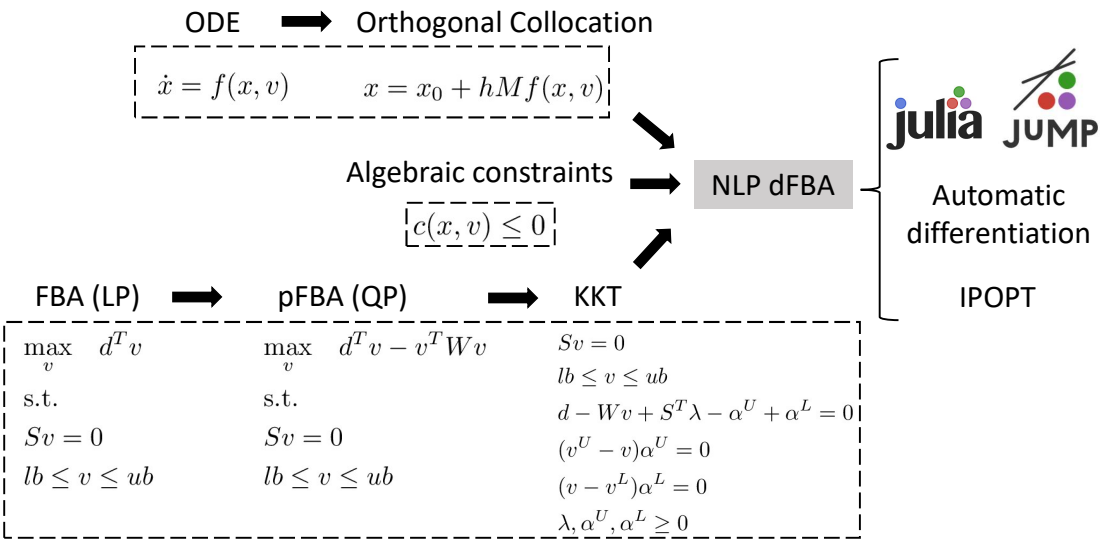


Figure 22 – Overview of the NLP dFBA approach. The ODE system is discretized using the orthogonal collocation technique on finite elements and inserted on the NLP dFBA as constraint. The algebraic constraints of dFBA related to uptake kinetics are directly inserted on the optimization problem as constraints. The FBA problem was first transformed into a pFBA by the addition of a quadratic regularization term. Then the KKT conditions of the pFBA were computed and inserted on the NLP dFBA as constraints. Finally the NLP dFBA was implemented in Julia language using the JUMP environment. An automatic differentiation package was used to compute the derivatives and the solver IPOPT was used to solve the resulting NLP problem.

4.2 Results

A series of case studies are presented in this section for different applications of dFBA models. First, the NLP dFBA is applied to the simple *E. coli* core model. The need for flux constraints and an adaptive mesh scheme is justified based on diauxic growth simulations. In the second case study the *E. coli* iJR904 model was applied, where the effect of the metabolic network size is explored.In the next case study the *E. coli* iJO1366 model was applied to estimate the optimal profile of genetic alterations for D-lactic acid production. Case study four evaluates the application of NLP dFBA for *S. cerevisiae* iND750 model in the dynamic optimization of a bioreactor problem. Finally, the NLP dFBA is applied to *S. cerevisiae* Yeast 8.3 model, which is the largest metabolic network applied in this work, solving a parameter estimation problem.

Study Case 1: *Escherichia coli* core model

The *E. coli* core model was chosen to be the first case study to test the methodology proposed because it is a simple and well-established model. First, a simulation of the aerobic growth on a medium containing glucose and acetate was performed. The model consists of mass balances for glucose (G), acetate (A), and biomass (X). Also, a Michaelis-Menten equation was applied for describing the glucose uptake:

$$\begin{aligned} \frac{dX}{dt} &= \mu X \\ \frac{dG}{dt} &= -v_g X \\ \frac{dA}{dt} &= v_A X \\ v_g &\leq v_{g,max} \frac{G}{K_g + G} \\ v_o &\leq v_{o,max} \\ X, G, A &\geq 0 \\ v_A,\ \mu &= pFBA(v_o, v_g) \end{aligned} \tag{4.6}$$

where μ, v_g, v_o, v_A are the growth rate, the glucose uptake flux, the oxygen uptake flux and the acetate flux, respectively. $K_g = 0.01\, mmol$ is the affinity constant for the substrate. $v_{g,max} = 10.5\, mmol/gdw\, h$ and $v_{o,max} = 19.0\, mmol/gdw\, h$ are the maximum uptake rate for glucose and oxygen, respectively. The acetate can be produced during aerobic growth on glucose or consumed when glucose is exhausted. It was considered that acetate can freely diffuse through the cell, and the acetate lower bound flux was set to $-2.5\, mmol/gdw\, h$.

The first simulation was performed until the moment of glucose exhaustion, for that case, no acetate consumption should occur. The NLP dFBA formulation presented in Equation 4.5 was used with an equidistant mesh of six elements and $\Phi(x, v) = 0$. In order to validate the solution obtained by the NLP formulation, the DFBAlab simulator was used as reference (GOMEZ et al., 2014). The lexicographic optimization order in DFBAlab was maximizing growth rate, minimizing glucose uptake, and maximizing acetate production. Also, the DA approach was applied using an explicit Runge-Kutta method from JULIA's DifferentialEquations package. The results showed a good agreement between the different approaches (Figure 23A).

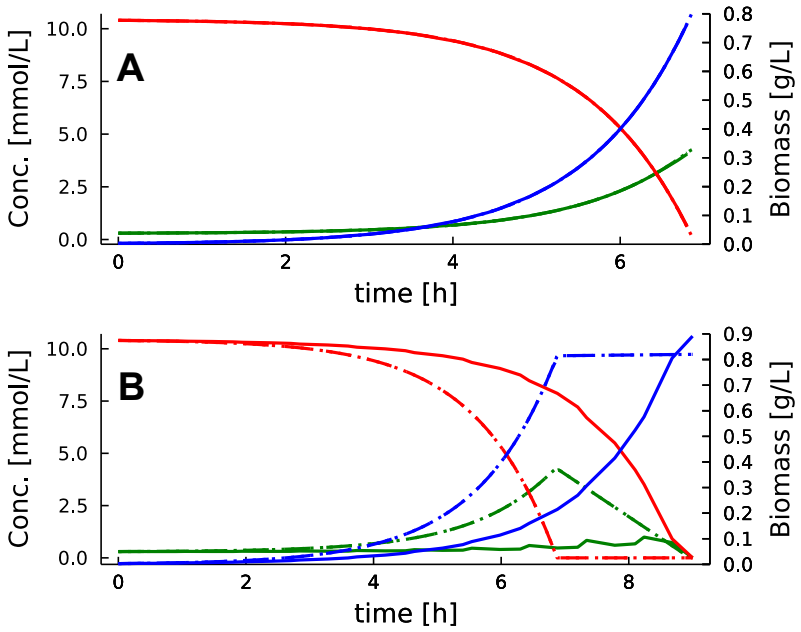


Figure 23 – *Escherichia coli* core model simulation until (a) glucose exhaustion and (b) acetate exhaustion. The concentration profiles are glucose (red), biomass (blue) and acetate (green). NLP dFBA (solid), DFBAlab (dotted) and direct approach (dashed). All the methods presented the same result for the *E. coli* growth on glucose (a), however, the NLP dFBA failed to describe the diauxic growth (b).

E. coli diauxic growth

A classic problem that can be simulated by dFBA models is the diauxic growth phenomena, when a microorganism in the presence of two substrates presents a two-phase growth, one for each substrate (MAHADEVAN et al., 2002). The dFBA should predict the right point when the preferential substrate is exhausted, and the cells start to consume the second substrate. In this example, *E. coli* should consume first the glucose and then acetate (MAHADEVAN et al., 2002). The NLP dFBA approach described the system dynamics poorly for this problem (Figure 23B). In order to diagnose the reasons for the NLP dFBA approach not to converge to the right solution, the acetate flux (v_{ac}) profile and the Lagrange multipliers associated to the acetate flux lower bound (α^L_{ac}) were examined more carefully. Table 7 shows the acetate flux and multipliers for each node in each finite element. As it can be seen, the flux of acetate alternates between the two possible optimal solutions, one consuming glucose and another consuming acetate. The last finite element is placed in the middle of the transition between the two phases. The multipliers for the last element are equal to zero for the first two nodes, and different from zero in the last one. This indicates that the acetate lower bound flux constraint becomes active within the element.

Biegler et al. (2002) noted that when there is a change of the active-set within a finite element, convergence problems can arise. They proposed an *ad hoc* method to deal with this problem. First, the NLP problem is solved and then it is evaluated if there is a change of active set within some finite element. Then, for these elements, the size is set as a variable. The methodology presented good results for optimal control applications, however, the need to solve the problem more than once makes the method not so practical to apply. Here, in order to deal with this problem, an extra constraint was imposed into the Equation 4.5 to prevent that flux distribution within a finite element from having a considerable change:

$$v_i^k - v_i^{k-1} \leq \epsilon \qquad k = 1, ..., ncp \tag{4.7}$$

where ncp is the number of collocation points, and ϵ is the allowed change in each flux. When this new constraint is added to the solution of the *E. coli* diauxic growth problem (using $\epsilon = 0$), the profiles that are obtained are closer to those from the other approaches (Figure 24A). Looking again at the acetate flux and multipliers profiles on the finite elements (Table 7), it is possible to see that there are no changes within a finite element. Moreover, only the last element is in the phase of consuming acetate.

Table 7 – Acetate flux and acetate lower bound multiplier profiles for each node in each finite element for the solution of *E. coli* diauxic growth using the NLP dFBA. When no flux constraint imposes the solution to jump between the two local solutions. However, when the flux constraints were imposed, there is no change in the flux and multipliers within an element.

	Element	1			2			3		
	Node	1	2	3	1	2	3	1	2	3
Without	α_{ac}^{L}	9.86e-9	9.86e-9	9.86e-9	9.86e-9	9.86e-9	9.86e-9	9.86e-9	9.86e-9	9.86e-9
flux const.	v_{ac}	4.13	-2.50	4.13	4.13	-2.50	4.13	4.13	-2.50	4.12
Applying	α_{ac}^{L}	9.62e-9	9.62e-9	9.62e-9	9.62e-9	9.62e-9	9.62e-9	9.62e-9	9.62e-9	9.62e-9
flux const.	v_{ac}	4.13	4.13	4.13	4.13	4.13	4.13	4.04	4.04	4.04

	Element	4			5			6		
	Node	1	2	3	1	2	3	1	2	3
Without	α_{ac}^{L}	9.86e-9	9.86e-9	9.86e-9	9.86e-9	9.86e-9	9.86e-9	9.86e-9	-0.02	-0.02
flux const.	v_{ac}	4.12	-2.50	4.12	4.12	-2.50	4.08	3.76	-2.50	-2.5
Applying	α_{ac}^{L}	9.59e-9	9.59e-9	9.59e-9	9.51e-9	9.51e-9	9.51e-9	-0.02	-0.02	-0.02
flux const.	v_{ac}	3.73	3.73	3.73	2.62	2.62	2.62	-2.50	-2.50	-2.50

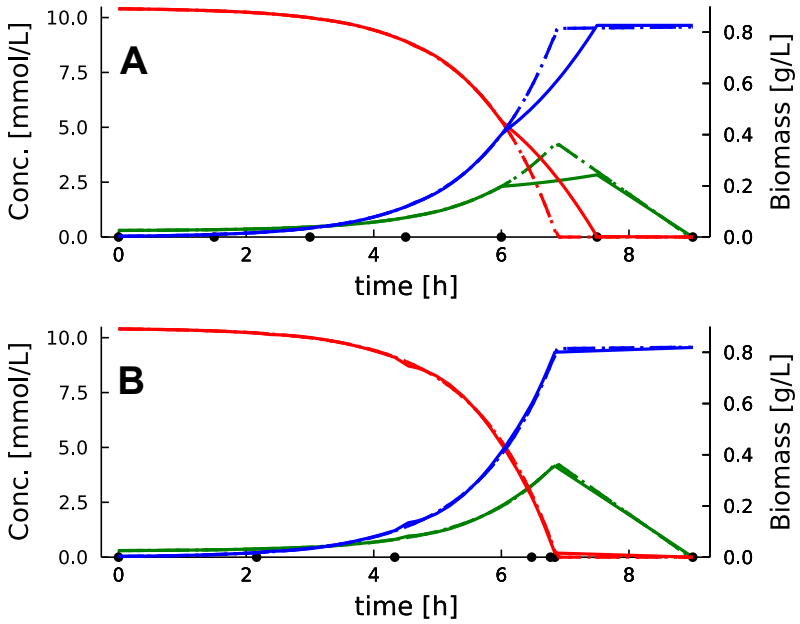


Figure 24 – *Escherichia coli* core model simulation until acetate exhaustion for NLP dFBA (-), DFBAlab (dotted) and direct approach (dashed). (A) imposing the constraint on the flux profile and (B) also using a adaptive mesh. RMSE [%] for glucose, acetate and biomass for DFBAlab/DA were 0.81/1.57, 0.70/1.07, 0.89/1.65, respectively. CPUs for NLP dFBA, DFBAlab and DA were 0.46, 0.99, 13.79 (seconds), respectively.

Adaptive mesh

Despite the profiles presenting a better agreement with other approaches when the flux constraints were imposed (Figure 24B), the transition between the two phases is still poorly described. This happens because the last element, which is in the consuming acetate phase, starts before the transition of the two phases. Therefore, to solve this issue, a simple adaptive mesh strategy was implemented as described in (BIEGLER, 2010):

$$\sum_{k=1}^{ne} h_k = t_{end}, \quad h_k \in [(1-\gamma)\overline{h}, (1+\gamma)\overline{h}] \tag{4.8}$$

In this strategy, the size of each element (h_k) is now a variable in the NLP and it is allowed to change in a range based on the equidistant element size ($\overline{h}$). For this problem, the element sizes were allowed to change freely ($\gamma = 1$). The profiles now have a good agreement with other approaches (Figure 24B). The position of the elements is indicated with circles on the bottom the Figure 24. The last element now begins exactly at the point of transition between phases, then the profile can be more accurately described. In order to

have a more precise measure of the difference between the approaches to solve the dFBA model, the Normalized Root Mean Square Error (RMSE) was calculated as follows:

$$RMSE[\%] = \frac{\sqrt{\frac{\sum_{i=1}^{N}(x_i^a - x_i^b)^2}{N}}}{max(x_i)} \tag{4.9}$$

where N is the number of samples, indices a and b correspond to different approaches and $max(x_i)$ is the maximum value of the metabolite. RMSE [%] for glucose, acetate and biomass for DFBAlab/DA were 0.81/1.57, 0.70/1.07, 0.89/1.65, respectively. CPU times for NLP dFBA, DFBAlab and DA were 0.46, 0.99, 13.79 (seconds), respectively. The adaptive mesh makes the problem more flexible with the cost of a bigger optimization problem to solve. However, by tuning the γ parameter this issue can be softened. For this small metabolic network, no considerable difference in the time to solve the problem was noted between NLP dFBA or DFBAlab. However, the DA method spends more time solving the problem. This can be explained by the need for the DA to solve more optimizations problems. Two for each step, because the bi-level version of pFBA was used.

Study Case 2: *E. coli* iJR904 model

The size of the metabolic network can have a considerable impact on the time and precision of dFBA simulations. Many authors suggested that an optimization approach would be intractable for genome-scale metabolic networks simulations (GOMEZ et al., 2014; SCOTT et al., 2018; OLIVEIRA et al., 2021a). Indeed, optimization approaches can suffer from scalability issues if the optimization problem becomes too big to be handled. Therefore, the *E. coli* iJR904 genome-scale model (GSM) (REED et al., 2003) was applied to simulate the diauxic growth problem in Equation 4.6. The *E. coli* iJR904 model has 761 metabolites and 1075 reactions, which is ten times bigger than the *E. coli* core model solved in the last section. The problem was solved using 9 adaptive elements. The results are shown in Figure 25, the RMSE [%] for glucose, acetate, and biomass for DFBAlab/DA were 0.85/4.88, 4.17/12.7, 0.79/4.48, respectively. CPU times for NLP dFBA, DFBAlab, and DA were 101.11, 1.05, and 204.62 seconds, respectively.

The DA method failed to predict the transition between the two substrate phases, thus explaining the higher RMSE deviation. Indeed, the NLP dFBA had a significant increment in computational time for the GSM while the DFBAlab had almost the same CPU times.

This can be explained by the fact that it is much easier to solve larger LP problems than NLP problems. Moreover, the fact that the DFBAlab computes an optimal basis for the FBA problem and avoids solving the optimization problem many times during the simulations explains the differences for the DA approach. It is important to highlight that despite the larger amount of time to solve the NLP dFBA, it is still a short time and also, the utilization of initial guess for the NLP problem can reduce that time (SHIN et al., 2019).

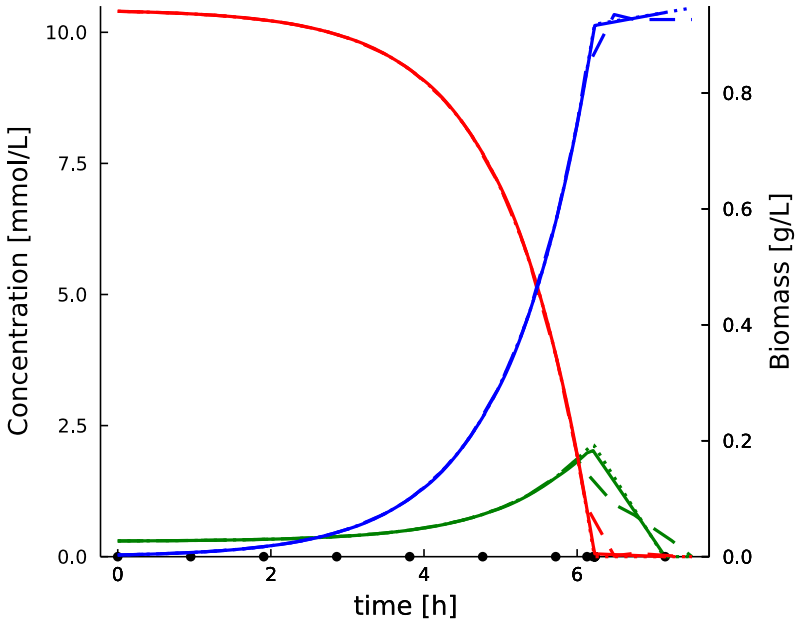


Figure 25 – *E. coli* iJR904 model diauxic growth simulation. The concentration profiles are glucose (red), biomass (blue) and acetate (green). NLP dFBA (solid), DFBAlab (dotted) and direct approach (dashed). RMSE [%] for glucose, acetate and biomass for DFBAlab/DA are 0.85/4.88, 4.17/12.7, 0.79/4.48, respectively. CPU seconds for NLP dFBA, DFBAlab and DA were 101.11, 1.05 and 204.62, respectively. The position of the finite elements is indicated with black circles.

Dynamic control of metabolism

An useful application of dFBA model is the dynamic control of metabolism. Gadkar et al. (2005) demonstrated that it is possible to apply dFBA models to estimate the optimal profile of metabolic fluxes to maximize the productivity of a bioproduct. The authors study the problem of maximizing ethanol production by *E. coli* regulating the flux through the acetate pathway. When acetate is being produced in anaerobiosis the cell can produce more ATP and have a high growth rate, on the other hand, when no acetate is being produced, the growth rate decreases but the ethanol production increases. The role of the optimization problem is to find the right time to switch from one state to another. The problem was

formulated as a bi-level optimization problem, in the upper level they maximize ethanol at the end of the batch by manipulating the time of regulation (t_{reg}). In the lower level, the dFBA model was solved using the direct approach. Here, the same problem was solved using the *E. coli* iJR904 metabolic network. The problem was solved by the NLP dFBA in a single-level optimization and the objective function in Equation 4.5 was set to $\Phi(x, v) = ethanol(end)$. Three finite elements were used, in the last two elements the constraint of a null flux on the acetate pathway were imposed. Because the size of the elements can change, the NLP can choose the t_{reg} based on the position of these elements.

Figure 26 shows the dynamic flux profile computed by NLP dFBA and the profiles for the wild-type strain and the static deletion strategy (when the acetate pathway is blocked since the beginning of the batch). As expected, the dynamic control strategy was able to increase ethanol production by manipulating the flux through acetate. Even though the direct comparison with the results presented in Gadkar et al. (2005) is not possible because here a GSM model was applied instead of a core model, the profiles are very similar.

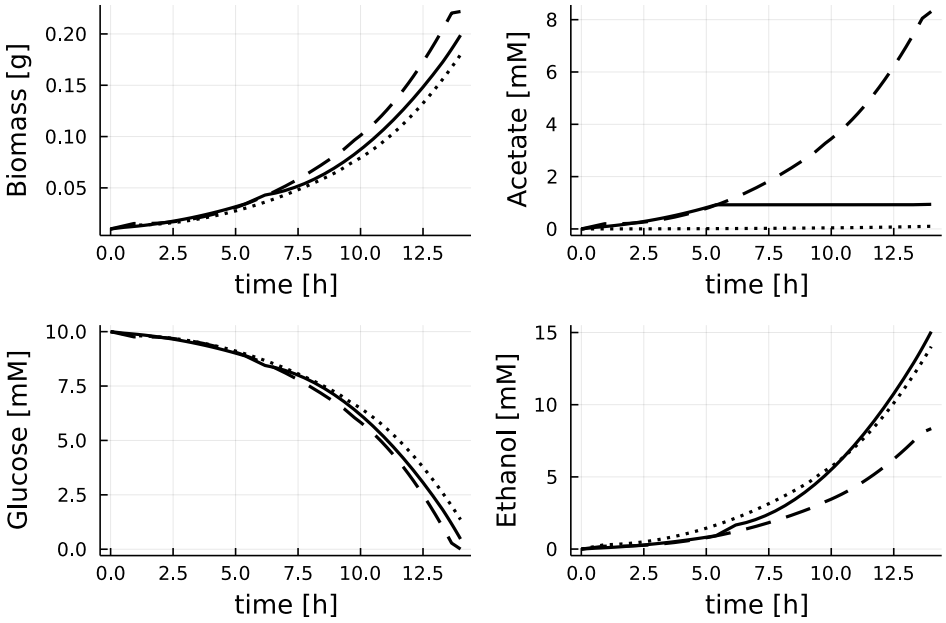


Figure 26 – Dynamic control of metabolism applying the *E. coli* iJR904 GSM solving by the NLP dFBA approach (solid). Static deletion simulation (dotted) and wild type simulation (dashed) are also presented.

This problem was also solved in DFBAlab as a bi-level optimization problem using the fmincon (SQP method) solver in MATLAB. In the outer optimization problem, t_{reg} was set as variable, and in the inner problem, two DFBAlab simulations were performed (one for each phenotype) based on t_{reg}. The formulation of this problem in DFBAlab illustrates

that the implementation is not so straightforward as the dFBA is already formulated as an optimization problem. The results of both methods were similar (Table 8) in terms of the time of switch and the CPUs time to solve the problem. Despite the fact that dFBAlab presented smaller time of simulation for the diauxic growth on *E. coli*, here, because the NLP dFBA is already formulated as an optimization problem, the computational cost to perform a simulation is almost the same as performing an optimal control simulation. This difference will be more relevant when solving larger problems as it will be evident in the next case studies.

Table 8 – Comparison of NLP dFBA and DFBAlab for solving the dynamic control of metabolism case study problem.

	DFBAlab	NLP dFBA
Number of elements	–	3
NLP iter	14	97
CPU time (s)	38.72	36.16
t_{reg} (h)	5.42	5.39
Ethanol (mM)	15.06	15.06

Case study 3: *E. coli* iJO1366 model

Despite the improvement of the two-stage process in the ethanol titer problem presented in the last section, the difference between the static strategy and the dynamic control strategy was small. Raj et al. (2020) demonstrated that for the two-stage process to overcome the one-stage process in productivity the second stage must have a non zero growth rate. Moreover, the glucose uptake profile can also have a significant impact on the analysis. Here, we reproduced the results presented in Raj et al. (2020) for the optimization of D-lactic acid productivity in *E. coli* using the NLP dFBA. The *E. coli* iJO1366 model was used (1805 metabolites and 2583 reactions) (ORTH et al., 2011). The dFBA model used as described by Raj et al. (2020) was:

$$
\begin{aligned}
\frac{dX}{dt} &= \mu X \\
\frac{dG}{dt} &= -v_g X \\
\frac{dL}{dt} &= v_L X \\
v_g &= v_{g,min} + (v_{g,max} - v_{g,min}) * (-1 + \frac{2}{(1 + exp(-K_{up} * \mu))}) \\
X, G, L &\geq 0 \\
v_L,\ \mu &= pFBA(v_g)
\end{aligned}
\tag{4.10}
$$

where μ, v_g, v_L are the growth rate, the glucose uptake flux and the D-lactic acid flux, respectively. $K_{up} = 5$ is the uptake constant for the glucose substrate. $v_{g,max} = 10.0\, mmol/gdw\, h$ and $v_{g,min} = 0.5\, mmol/gdw\, h$ are the maximum and minimum uptake rate for glucose, respectively. The model was solved using the logistic curve to describe the glucose uptake (reduced) as in Equation 4.10 and using a constant glucose uptake rate equal to $v_{g,max}$. Raj et al. (2020) formulated a bi-level optimization problem to maximize the D-lactic acid productivity in a two-stage process. Here, the problem was solved by the NLP dFBA in a single-level optimization and the objective function in Equation 4.5 was set to $\Phi(x, v) =$ D-lactic acid(end)$/h$. Six finite elements were used with a constraint that the first three elements must have the same growth rate, and the same for the last three elements. Because the size of the elements can change, the NLP can choose the t_{reg} based on the position of these elements.

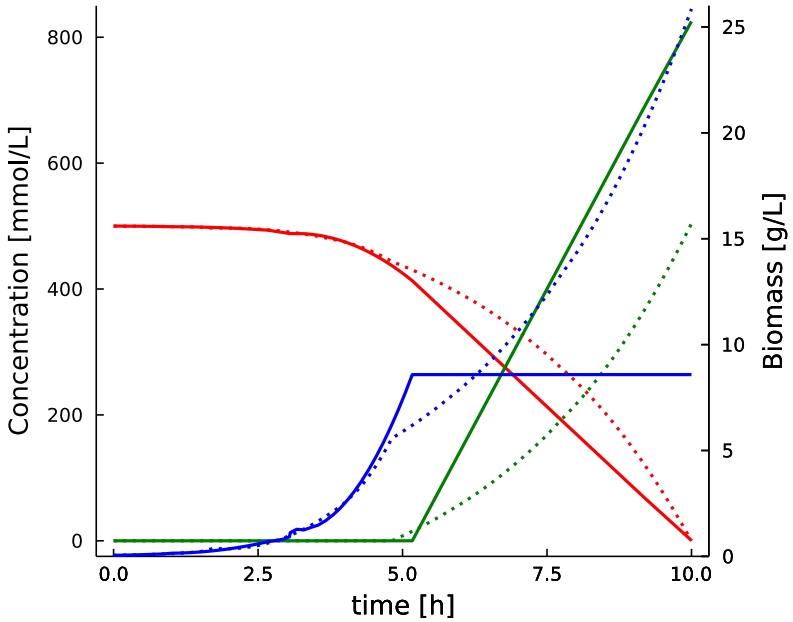


Figure 27 – D-lactic acid production in *E. coli* iJO1366 model for constant glucose uptake (solid line) and reduced glucose uptake (dotted line) using the NLP dFBA. When the glucose uptake is used the second stage has a null growth rate and the maximum D-lactic acid production. However, when the reduced uptake is applied, the second stage must have a non zero growth rate.

Figure 27 show the profiles for both cases (constant and reduced glucose uptake). As already described by Raj et al. (2020),when the glucose uptake is held at its maximum value, there is a first stage of growth only and a second stage of D-lactic acid production only. On the other hand, when the effects in the glucose uptake are considered, the best profiles change, and the second optimum stage has a growth rate that decreases the D-lactic acid production. Table 9 shows the change in the stages when glucose uptake is constant and reduced. Maximum growth rate is 0.9 and maximum v_L is 20. t_{reg} was similar to the one reported by Raj et al. (2020) demonstrating the capacity of NLP dFBA of reproducing the optimal solution using a single-level approach.

Table 9 – Process parameters for the D-lactic acid production in *E. coli* iJO1366 model using the NLP dFBA.

	Glucose uptake	
	constant	reduced
$\mu\ (1/h)$	0.9 ⇒ 0.0	0.9 ⇒ 0.3
$v_L\ (mmol/gdw\,h)$	0.0 ⇒ 20	0.0 ⇒ 7.35
$t_{reg}\ (h)$	5.17	4.81
Productivity $(mmollac/gdw\,h)$	82.52	50.41
Yield $(mmol\,lac/mmol\,glu)$	1.65	1.01
Titer $(mmol\,lac/gdw)$	825	504

Study Case 4: *Saccharomyces cerevisiae* iND750

Ethanol production by *S. cerevisiae* is still an active subject of research, both in terms of genetic manipulations and bioreactor control and optimization. dFBA model can be useful in both cases and this method was applied to simulate the metabolism of *S. cerevisiae* using the GSM iND750 model containing 1061 metabolites and 1266 reactions (DUARTE et al., 2004). A bioreactor model for the production of ethanol by controlling the glucose feed profile and the dissolved oxygen level was formulated based on previous work (CHANG et al., 2016; OLIVEIRA et al., 2021a):

$$
\begin{aligned}
\frac{dV}{dt} &= F \\
\frac{d(VX)}{dt} &= \mu VX \\
\frac{d(VG)}{dt} &= FG_f - v_g VX \\
\frac{d(VE)}{dt} &= v_e VX \\
v_g &\leq v_{g,max} \frac{G}{K_g + G + (G^2/K_{ig})} \frac{1}{1 + (E/K_{ie})} \\
v_o &\leq v_{o,max} \frac{O}{K_o + O} \\
X, G, E &\geq 0 \\
v_e,\ \mu &= pFBA(v_o, v_g)
\end{aligned}
\tag{4.11}
$$

where μ, v_g, v_o, v_e are the growth rate, the glucose uptake flux, the oxygen uptake flux and the ethanol flux, respectively. $K_g = 0.5\, g/L$, $K_{ig} = 10\, g/L$, $K_{ie} = 10.0\, g/L$ and K_o = 3.0x 10^{-6} mol/L are the glucose saturation constant, glucose inhibition constant, ethanol Inhibition constant and oxygen saturation constant, respectively. $v_{g,max} = 20.0\, mmol/gdw\, h$ and $v_{o,max} = 8.0\, mmol/gdw\, h$ are the maximum uptake rate for glucose and oxygen, respectively. G_f is the feed glucose concentration and was fixed in $50.0\, g/L$. The control variables were the glucose feed (F) and oxygen concentration that was assumed to be direct manipulated by a perfect feedback controller. The percentage of dissolved oxygen is calculated as a function of the oxygen saturation concentration ($DO = O/O_{sat}$) and O_{sat} was set to 3.0x 10^{-4} mol/L.

First, a simple simulation was performed in order to compare the performance of the different approaches. Figure 28 present the results, the RMSE [%] for biomass, glucose, and ethanol for DFBAlab/DA were 0.41/1.75, 0.42/2.35, 0.62/1.44, respectively. The CPUs for NLP dFBA, DFBAlab, and DA were 203.37, 1.38, 31.85, respectively. As can be seen again,

a good agreement between the NLP dFBA and dFBAlab was achieved. However, the DA approach failed to represent the profiles after the step in the control variables. NLP dFBA had a considerable increase of time of simulation in comparison with the last examples.

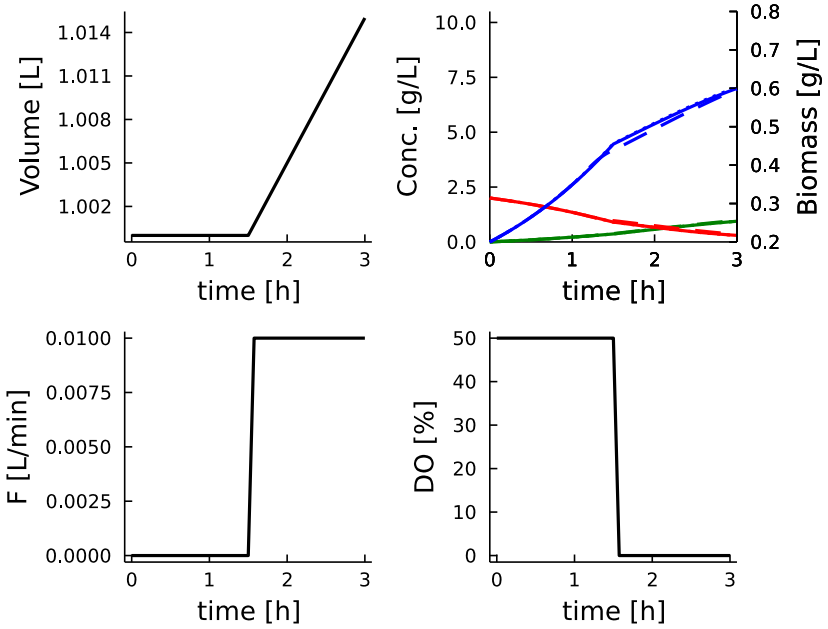


Figure 28 – Bioreactor model simulation using *Saccharomyces cerevisiae* iND750 model. RMSE [%] for biomass, glucose, and ethanol for DFBAlab/DA were 0.41/1.75, 0.42/2.35, 0.62/1.44, respectively. CPUs for NLP dFBA, DFBAlab and DA were 203.37, 1.38, 31.85, respectively.

Dynamic optimization of bioreactors

Besides the application of the dFBA model to dynamic control of the metabolism, this model can also be applied to dynamic optimization/optimal control of a bioprocess. Chang et al. (2016) formulated a model predictive control using a KKT reformulation of the FBA problem. The authors reported convergence and initialization problems when trying to use a GSM and consequently they were forced to apply a core model of *S. cerevisiae*. Oliveira et al. (2021a) was able to apply a GSM for the same problem, however, they needed to replace the FBA optimization problem by a surrogate model, adding a training step to the process. Scott et al. (2018) formulated two dynamic optimization problems for *E. coli* and *Pichia stipitis* using GSM. Scott et al. (2018) solved the dFBA as an implicit ODE system and computed the gradient information by solving the sensitivity equation.

The model presented in Equation 4.11 was applied to dynamic optimization of the ethanol production in the bioreactor. The aim is to control the glucose feed flow rate to avoid

substrate inhibition avoiding the violation of the maximum volume constraint of the bioreactor (1.2 L). The DO inside the reactor must be manipulated to switch from a growth phase (high DO) to a production phase (low DO). The NLP dFBA model was formulated using the objective function in Equation 4.5 as $\Phi(x, v) = ethanol(end)$ and the process constraints were also imposed to the problem ($V_{final} \leq 1.2$ and $u_{min} \leq u \leq u_{max}$). The problem was also solved using DFBAlab+fmincon as a bi-level optimization problem. The control variables were discretized into 10 elements for both approaches.

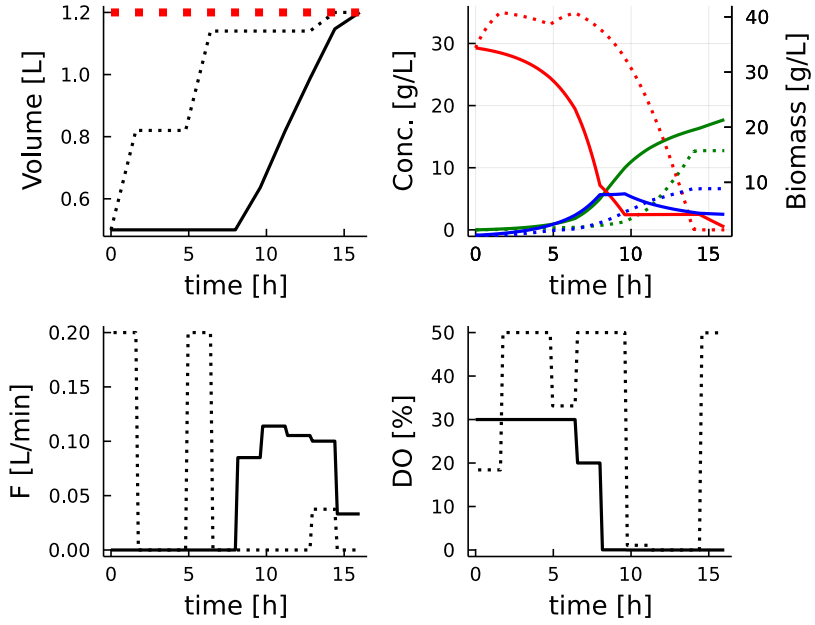


Figure 29 – Dynamic optimization simulation of a bioreactor using the *Saccharomyces cerevisiae* iND750 model using NLP dFBA (solid) and DFBAlab (dotted). The concentration profiles are glucose (red), biomass (blue) and ethanol (green).

The results are presented in Figure 29 and Table 10. The NLP dFBA model reached a solution with better performance in comparison to the solution found using DFBAlab in terms of ethanol final mass. The NLP dFBA profiles have two clearly distinct phases (growth and production) which is analogous to the profiles obtained by Chang et al. (2016) and Oliveira et al. (2021a) in recent studies. On the other hand, the DFBAlab profiles were less smooth and because of a higher feed flow at the beginning of the batch, glucose inhibition effects takes place. Looking at the numerical aspects, the NLP dFBA takes almost the same time to perform a simulation or optimization. This is expected because of the optimization nature of the problem. DFBAlab took a long time to converge using fmincon mostly because of the lack of gradient information, that must be estimated by finite differences, and also because of the presence of nonlinear constraints ($V_{final} \leq 1.2$). Another problem with DFBAlab for this

application was pointed out by Scott et al. (2018), is the presence of discontinuities in the model at the points where events take place.

It is important to mention that the methodology presented by Scott et al. (2018) represents an alternative for supplying gradient information of dFBA models. Nonetheless, the transcription of the dynamic model into algebraic equation allows the utilization of automatic differentiation packages that makes the solution of optimization problems more straightforward. In contrast, simulation-based computation of derivatives information (e.g. finite difference forward sensitivities) are computationally expensive and differential equation solvers can fail at some perturbation inputs making the process unstable (SHIN et al., 2019).

Table 10 – Comparison of the solution of the dynamic optimization of bioreactor using NLP dFBA and DFBAlab.

	DFBAlab	NLP dFBA
number of elements	10	10
function evaluations	195	268
NLP iterations	6	174
CPU time (min)	40.15	3.27
Ethanol (g)	15.29	21.282

Case study 5: *Saccharomyces cerevisiae* Yeast 8.3 model

In spite of the fact that dFBA models have fewer parameters than detailed kinetic models, there are still some parameters that need to be estimated using experimental data. As in dynamic control optimization, parameter estimation using dFBA becomes a bi-level problem and the non-smoothness of dFBA makes the problem hard to solve. Raghunathan et al. (2006) represent the dFBA model inside the estimation problem using a system of Differential Variational Inequalities that is discretized to yield a MPCC problem and then solved using a interior point algorithm. They applied the approach to a small-scale metabolic network of 39 reactions and 43 metabolites. Leppävuori et al. (2011) formulated a sequential gradient-based solution with direct sensitivity equations. They estimated 8 parameters and used a metabolic network of 1266 enzymatic reactions and 1061 metabolites. Waldherr (2016) solved a parameter estimation problem for a small-scale network of 10 reactions and 12 metabolites. The bi-level problem was solved as an mixed integer quadratic program. Oliveira et al. (2022) solved a parameter estimation of 5 parameters using a Yeast metabolic

network of 2666 metabolites and 3928 enzymatic reactions. They trained a surrogate model to replace the FBA optimization inside the dFBA model.

Here, we applied NLP dFBA to solve the estimation problem formulated by Oliveira et al. (2022) using the Yeast 8.3 metabolic network model (HEAVNER et al., 2013). The dFBA model describes the *S. cerevisiae* anaerobic growth on glucose and xylose substrates (Equation 4.12).

$$\begin{aligned}
\frac{dX}{dt} &= \mu X \\
\frac{dG}{dt} &= -v_g X \\
\frac{dZ}{dt} &= -v_z X \\
\frac{dE}{dt} &= v_e X \\
v_g &\leq v_{g,max} \frac{G}{K_g + G} \\
v_z &\leq v_{z,max} \frac{Z}{K_z + Z} \frac{1}{1 + (G/K_{ig})} \\
X, G, E, Z &\geq 0 \\
v_e,\ \mu &= pFBA(v_g, v_z)
\end{aligned} \tag{4.12}$$

where μ, v_g, v_z, and v_e are the growth rate, and the exchange fluxes of glucose, xylose, and ethanol, respectively. X, G, Z, and E represent the biomass, glucose, xylose, and ethanol concentrations, respectively. $v_{g,max}$ and $v_{z,max}$ are the maximum uptake rate for glucose and xylose respectively. K_g and K_z are the saturation constants, and K_{ie} is the glucose inhibition constant.

A dFBA simulation using the direct method in MATLAB was applied to yield the *in silico* data of glucose, xylose, biomass and ethanol. The set of parameters used are presented in Table 11. The parameter estimation problem was implemented as an NLP dFBA using the objective function in Equation 4.5 as $\Phi(x, v) = \sum_j (x_j^c(\theta) - x_j^m)^2$. Indices c and m indicate calculated and measured quantities respectively. Multiple initial guesses were supplied to the solver and the best fit parameters simulation is presented in Figure 30. The NLP dFBA method was able to find a good fit to the data and the parameters were similar to the ones used to generate the *in silico* data. The features of the performance of NLP dFBA can also be found in Table 11. 12 finite elements were used to solve the problem culminating in a large optimization problem to solve. Considering that the problem was solved in a personal laptop, the CPU time was adequate. The uncertainty quantification of the estimated

parameters is a crucial step on the model identification process, some techniques rely on solving repeated parameter estimation problems (SHIN et al., 2019). Therefore, a fast and reliable method as NLP dFBA is needed.

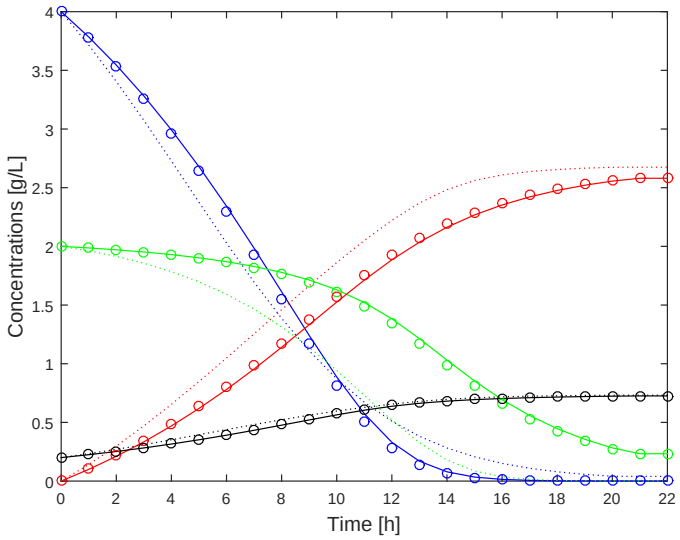


Figure 30 – Best fit of dFBA using *S. cerevisiae* Yeast 8.3 model to the *in silico* data points (circles) computed by NLP dFBA (solid line) and the direct approach (dotted line). The concentration profiles are glucose (red), biomass (blue) and acetate (green).

Table 11 – Computational Performance of NLP dFBA to solve the parameter estimation problem and the parameter values used in model simulation to yield measurements.

Parameters	Model simulation	NLP dFBA	Direct approach
v_g^{max}	7.30	7.11	30.01
K_g	1.03	1.01	12.02
v_z^{max}	32.00	32.88	7.99
K_z	14.85	15.16	0.80
K_{ie}	0.50	0.48	1.00
number of elements	-	12	-
variables	-	268026	-
constraints	-	362496	-
CPU	-	26.02 min	27.82 min
iterations	-	628	5
function evaluations	-	1440	36
Objective function	-	1.89e-3	7.17

Sequential approach

Finally, we want to highlight another possible approach to solve dFBA models using NLP dFBA. In addition to the single optimization approach that we have been using so far in this work, another possibility is to use a sequential approach as illustrated in Figure 31. On the sequential approach, the time domain is discretized and one optimization problem is solved for each interval. This approach improves the CPU time for big problems (Table 12) and also deals well with the problems pointed out in Section 4.2 on the diauxic growth simulations. This approach is similar to the one used by Scott et al. (2018) with the distinction that each step is solved by an optimization problem using an orthogonal collocation method instead of solving as an implicit ODE system. This approach can be mainly useful to perform simulations in a short time or to solve large systems (e.g. human cells GSM) that otherwise would culminate in a large single optimization problem. However, for small/medium size GSM simulation problems, dynamic optimization, or parameter estimation problems, the single optimization approach still being the best approach because the gradients can be computed simultaneously. To solve this issue, in the future, techniques such as the quasi-sequential approach could be explored (HONG et al., 2006).

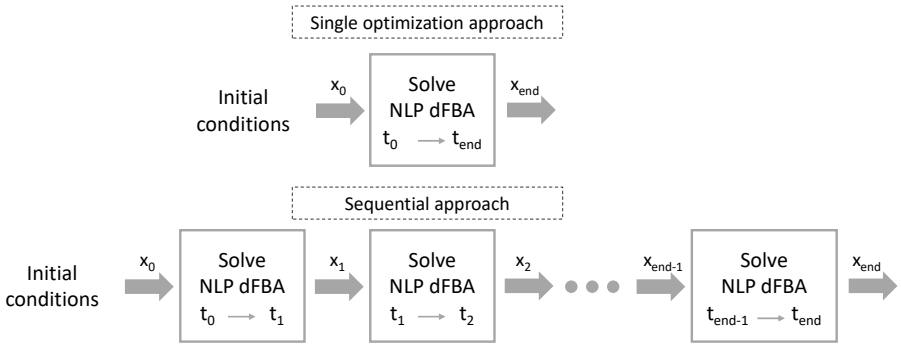


Figure 31 – Comparison between the sequential and single optimization approaches for the NLP dFBA.

Table 12 – CPUs time comparison for simulations with different metabolic network models.

	DFBAlab	NLP dFBA single opt.	NLP dFBA sequential
E. coli core	0.99	0.46	1.11
E. coli iJR904	1.05	101.11	12.34
S. cerevisiae iND750	1.38	31.85	6.91

4.3 Conclusions

An NLP formulation of dFBA model was presented (NLP dFBA). The method consists in reformulating the pFBA using the KKT conditions and discretizing the dynamic model using the orthogonal collocation technique. The method was implemented in the high-performance JULIA language and solved by the large-scale NLP solver IPOPT. An automatic differentiation code package was applied to supply first and second-order derivative information. NLP dFBA represents a new attempt to solve dFBA models as an optimization problem. As demonstrated, this approach can significantly improve the solution of a class of problems in comparison with the simulation approach. We hope that the tool developed here will contribute to fields such as metabolic engineering, synthetic biology, and bioprocess optimization. From the study cases of simulation of *E. coli* diauxic growth on glucose and acetate the need to impose some flux constraints to guarantee the convergence of the method when there is a change in the active-set of the FBA optimization problem was evidenced. Furthermore, the advantage of using an adaptive mesh strategy was also highlighted. The NLP dFBA was applied to different metabolic networks and scaled well for genome-scale networks. NLP dFBA is a more straightforward way to solve optimization problems with embedded dFBA models. This was demonstrated when the method was applied to simulate a dynamic control of metabolism and to solve a dynamic optimization of a bioreactor and a parameter estimation problem.

In future work, some extensions of the methodology and some new applications may be explored, like a quasi-sequential approach that can further decrease the CPU time (HONG et al., 2006). The methodology developed here can also be extended to problems beyond dFBA models, such as kinetic dynamic models that apply an optimality principle to obtain a solution (TSIANTIS et al., 2018).

5 Future work

A recommendation of work for the identifiability analysis of metabolic flux ratios on carbon labeling experiments is to take into consideration the nonlinearity of the flux space for the methods that apply Principal Component Analysis and the Fisher Information Matrix. The labeling pattern of the substrates determined by the Design of Optimal Labeling Experiments can be locally optimal but not globally optimal. This effect was illustrated when the exchange flux on the Pentose phosphate pathway goes from low values to close to the equilibrium. A sampling technique can be applied to explore the whole flux space. Furthermore, the application of thermodynamic constraints and Flux Balance Analysis for reducing the searching space can also be explored. Another possible extension of the work would be to the isotopic non-stationary condition (^{13}C-NMFA), where more information about the flux distribution can be obtained. However, a system of nonlinear differential equations must be solved in the identifiability analysis and the number o variables are higher because the pool sizes of the metabolites must be taken into account (WIECHERT; NöH, 2013). These characteristics makes DOE harder to solve.

An interesting work on the surrogate FBA method would be to develop an approach to train automatically the surrogate model from the Flux Balance Analysis simulations. This method could be used to determine the amount of sampling points to obtain, but also the type of surrogate models to be applied. In the case of piecewise polynomial models, the regions should also be determined. The surrogate FBA has a good potential to scale to larger networks, for example those that include protein synthesis networks. This makes the application of these models suitable to integrate inside a whole-cell model (ELSEMMAN et al., 2022). Finally, process synthesis design can take advantages of dFBA simulations, and the surrogate FBA can be appropriate for a superstructure optimization (AKBARI; BARTON, 2019).

Applying the NLP dFBA approach to large metabolic network models (e.g. Human metabolic network model) would be interesting and challenging to solve, new modifications to the NLP dFBA methodology will be required, a quasi-sequential approach can be explored (HONG et al., 2006).

Bibliography

AKBARI, A.; BARTON, P. I. Integrating Genome-Scale and Superstructure Optimization Models in Techno-Economic Studies of Biorefineries. *Processes*, v. 7, n. 5, maio 2019. ISSN 2227-9717. Number: 5 Publisher: Multidisciplinary Digital Publishing Institute.

ANESIADIS, N.; CLUETT, W. R.; MAHADEVAN, R. Dynamic metabolic engineering for increasing bioprocess productivity. *Metabolic Engineering*, v. 10, n. 5, p. 255–266, 2008. ISSN 1096-7176.

ANTONIEWICZ, M. R. 13c metabolic flux analysis: optimal design of isotopic labeling experiments. *Current Opinion in Biotechnology*, v. 24, n. 6, p. 1116–1121, dez. 2013. ISSN 09581669.

ANTONIEWICZ, M. R. Methods and advances in metabolic flux analysis: a mini-review. *Journal of Industrial Microbiology & Biotechnology*, v. 42, n. 3, p. 317–325, mar. 2015. ISSN 1367-5435, 1476-5535.

ANTONIEWICZ, M. R.; KELLEHER, J. K.; STEPHANOPOULOS, G. Elementary metabolite units (EMU): A novel framework for modeling isotopic distributions. *Metabolic Engineering*, v. 9, n. 1, p. 68–86, jan. 2007. ISSN 10967176.

ASTROM, K. J.; MURRAY, R. M. *Feedback Systems: An Introduction for Scientists and Engineers*. USA: Princeton University Press, 2008. ISBN 0691135762.

BARD, Y. *Nonlinear Parameter Estimation*. [S.l.: s.n.], 1974.

BAUMRUCKER, B. T.; RENFRO, J. G.; BIEGLER, L. T. MPEC problem formulations and solution strategies with chemical engineering applications. *Computers & Chemical Engineering*, v. 32, n. 12, p. 2903–2913, dez. 2008. ISSN 0098-1354.

BEERS, K. J. *Numerical methods for chemical engineering: applications in Matlab*. Cambridge; New York: Cambridge University Press, 2007. OCLC: 708245930. ISBN 978-0-511-25538-0 978-0-511-64881-6 978-0-511-25483-3 978-0-511-25593-9 978-0-511-25650-9 978-0-511-81219-4.

BERGER, A. et al. Robustness and Plasticity of Metabolic Pathway Flux among Uropathogenic Isolates of Pseudomonas aeruginosa. *PLoS One*, v. 9, n. 4, abr. 2014. ISSN 1932-6203.

BEZANSON, J. et al. Julia: A fresh approach to numerical computing. *SIAM review*, SIAM, v. 59, n. 1, p. 65–98, 2017.

BIEGLER, L. T. *Nonlinear Programming*. [S.l.]: Society for Industrial and Applied Mathematics, 2010.

BIEGLER, L. T.; CERVANTES, A. M.; WäCHTER, A. Advances in simultaneous strategies for dynamic process optimization. *Chemical Engineering Science*, v. 57, n. 4, p. 575–593, fev. 2002. ISSN 0009-2509.

BRUNNER, J. D.; CHIA, N. Minimizing the number of optimizations for efficient community dynamic flux balance analysis. *PLOS Computational Biology*, v. 16, n. 9, p. e1007786, set. 2020. ISSN 1553-7358. Publisher: Public Library of Science.

BURG, J. M. et al. Large-scale bioprocess competitiveness: the potential of dynamic metabolic control in two-stage fermentations. *Current Opinion in Chemical Engineering*, v. 14, p. 121–136, 2016. ISSN 2211-3398. Biotechnology and bioprocess engineering / Process systems engineering.

CARDINALI-REZENDE, J. et al. The relevance of enzyme specificity for coenzymes and the presence of 6-phosphogluconate dehydrogenase for polyhydroxyalkanoates production in the metabolism of Pseudomonas sp. LFM046. *International Journal of Biological Macromolecules*, v. 163, p. 240–250, nov. 2020. ISSN 0141-8130.

CHANG, L.; LIU, X.; HENSON, M. A. Nonlinear model predictive control of fed-batch fermentations using dynamic flux balance models. *Journal of Process Control*, v. 42, p. 137–149, jun. 2016. ISSN 0959-1524.

CHANG, Y.; SUTHERS, P. F.; MARANAS, C. D. Identification of optimal measurement sets for complete flux elucidation in metabolic flux analysis experiments. *Biotechnology and Bioengineering*, v. 100, n. 6, p. 1039–1049, 2008. ISSN 1097-0290.

CHINDELEVITCH, L. et al. An exact arithmetic toolbox for a consistent and reproducible structural analysis of metabolic network models. v. 5, n. 1, p. 1–9, 2014. ISSN 2041-1723.

CHINDELEVITCH, L. et al. Reply to “do genome-scale models need exact solvers or clearer standards?”. v. 11, n. 10, 2015. ISSN 1744-4292.

CHOI, M. H. et al. Metabolic relationship between polyhydroxyalkanoic acid and rhamnolipid synthesis in Pseudomonas aeruginosa: Comparative 13c NMR analysis of the products in wild-type and mutants. *Journal of Biotechnology*, v. 151, n. 1, p. 30–42, jan. 2011. ISSN 0168-1656.

DUARTE, N. C.; HERRGåRD, M. J.; PALSSON, B. Reconstruction and Validation of Saccharomyces cerevisiae iND750, a Fully Compartmentalized Genome-Scale Metabolic Model. *Genome Research*, v. 14, n. 7, p. 1298–1309, jul. 2004. ISSN 1088-9051.

DUARTE, N. C.; PALSSON, B. ; FU, P. Integrated analysis of metabolic phenotypes in saccharomyces cerevisiae. v. 5, n. 1, p. 63, 2021. ISSN 1471-2164.

EBRAHIM, A. et al. Do genome-scale models need exact solvers or clearer standards? v. 11, n. 10, 2015. ISSN 1744-4292.

ELSEMMAN, I. E. et al. Whole-cell modeling in yeast predicts compartment-specific proteome constraints that drive metabolic strategies. v. 13, n. 1, p. 801, 2022. ISSN 2041-1723. Number: 1 Publisher: Nature Publishing Group.

FANG, X.; LLOYD, C. J.; PALSSON, B. O. Reconstructing organisms in silico: genome-scale models and their emerging applications. v. 18, n. 12, p. 731–743, 2020. ISSN 1740-1534. Number: 12 Publisher: Nature Publishing Group.

FOLLSTAD, B. D.; STEPHANOPOULOS, G. Effect of reversible reactions on isotope label redistribution. *European Journal of Biochemistry*, v. 252, n. 3, p. 360–371, 1997. ISSN 1432-1033.

FRITZEMEIER, C. J. et al. Erroneous energy-generating cycles in published genome scale metabolic networks: Identification and removal. *PLOS Computational Biology*, v. 13, n. 4, p. e1005494, abr. 2017. ISSN 1553-7358. Publisher: Public Library of Science.

GADKAR, K. G. et al. Estimating optimal profiles of genetic alterations using constraint-based models. *Biotechnology and Bioengineering*, v. 89, n. 2, p. 243–251, 2005. ISSN 1097-0290.

GELADI, P.; KOWALSKI, B. R. Partial least-squares regression: a tutorial. *Analytica Chimica Acta*, v. 185, p. 1–17, 1986. ISSN 00032670.

GOMEZ, J. A.; HöFFNER, K.; BARTON, P. I. DFBAlab: a fast and reliable MATLAB code for dynamic flux balance analysis. *BMC Bioinformatics*, v. 15, n. 1, p. 409, dez. 2014. ISSN 1471-2105.

GRIMSTAD, B. et al. Global optimization of multiphase flow networks using spline surrogate models. *Computers Chemical Engineering*, v. 84, p. 237–254, 2016. ISSN 0098-1354.

GUEBILA, M. B.; THIELE, I. Dynamic flux balance analysis of whole-body metabolism for type 1 diabetes. *Nature Computational Science*, v. 1, n. 5, p. 348–361, maio 2021. ISSN 2662-8457. Number: 5 Publisher: Nature Publishing Group.

GUSTAVSSON, M.; LEE, S. Y. Prospects of microbial cell factories developed through systems metabolic engineering. *Microbial Biotechnology*, v. 9, n. 5, p. 610–617, 2016. ISSN 1751-7915.

HARWOOD, S. M.; HöFFNER, K.; BARTON, P. I. Efficient solution of ordinary differential equations with a parametric lexicographic linear program embedded. *Numerische Mathematik*, v. 133, n. 4, p. 623–653, ago. 2016. ISSN 0029-599X, 0945-3245.

HEAVNER, B. D. et al. Version 6 of the consensus yeast metabolic network refines biochemical coverage and improves model performance. *Database (Oxford)*, v. 2013, ago. 2013. ISSN 1758-0463.

HJERSTED, J. L.; HENSON, M. A.; MAHADEVAN, R. Genome-scale analysis of Saccharomyces cerevisiae metabolism and ethanol production in fed-batch culture. *Biotechnology and Bioengineering*, v. 97, n. 5, p. 1190–1204, ago. 2007. ISSN 1097-0290.

HODGSON, B. J. et al. Intelligent modelling of bioprocesses: a comparison of structured and unstructured approaches. *Bioprocess Biosyst Eng*, v. 26, n. 6, p. 353–359, dez. 2004. ISSN 1615-7605.

HONG, W. et al. A quasi-sequential approach to large-scale dynamic optimization problems. *AIChE Journal*, v. 52, n. 1, p. 255–268, 2006. ISSN 1547-5905.

HORI, K.; MARSUDI, S.; UNNO, H. Simultaneous production of polyhydroxyalkanoates and rhamnolipids by Pseudomonas aeruginosa. *Biotechnology and Bioengineering*, v. 78, n. 6, p. 699–707, 2002. ISSN 1097-0290.

ISERMANN, N.; WIECHERT, W. Metabolic isotopomer labeling systems. Part II: structural flux identifiability analysis. *Mathematical Biosciences*, v. 183, n. 2, p. 175–214, jun. 2003. ISSN 00255564.

JABARIVELISDEH, B. et al. Adaptive predictive control of bioprocesses with constraint-based modeling and estimation. *Computers Chemical Engineering*, v. 135, p. 106744, 2020. ISSN 0098-1354.

KAPPELMANN, J.; WIECHERT, W.; NOACK, S. Cutting the Gordian Knot: Identifiability of anaplerotic reactions in Corynebacterium glutamicum by means of 13C-metabolic flux analysis. *Biotechnology and Bioengineering*, v. 113, n. 3, p. 661–674, mar. 2016. ISSN 1097-0290.

KITANO, H. Systems biology: a brief overview. *Science*, v. 295, n. 5560, p. 1662–1664, 2002.

KOHLSTEDT, M.; WITTMANN, C. GC-MS-based 13c metabolic flux analysis resolves the cyclic glucose metabolism of Pseudomonas putida KT2440 and Pseudomonas aeruginosa PAO1. *Metabolic Engineering*, mar. 2019. ISSN 1096-7176.

KUMAR, D.; BUDMAN, H. Robust nonlinear predictive control for a bioreactor based on a Dynamic Metabolic Flux Balance model. *IFAC-PapersOnLine*, v. 48, n. 8, p. 930–935, jan. 2015. ISSN 2405-8963.

KUMAR, D.; BUDMAN, H. Applications of Polynomial Chaos Expansions in optimization and control of bioreactors based on dynamic metabolic flux balance models. *Chemical Engineering Science*, v. 167, p. 18–28, ago. 2017. ISSN 0009-2509.

KUYPER, M. et al. Minimal metabolic engineering of Saccharomyces cerevisiae for efficient anaerobic xylose fermentation: a proof of principle. *FEMS Yeast Res.*, v. 4, n. 6, p. 655–664, mar. 2004. ISSN 1567-1356.

LEE, E. Y.; CHOI, C. Y. Gas chromatography-mass spectrometric analysis and its application to a screening procedure for novel bacterial polyhydroxyalkanoic acids containing long chain saturated and unsaturated monomers. *Journal of Fermentation and Bioengineering*, v. 80, n. 4, p. 408 – 414, 1995. ISSN 0922-338X.

LEE, J. H. Model predictive control: Review of the three decades of development. *International Journal of Control, Automation and Systems*, Springer, v. 9, n. 3, p. 415, 2011.

LEE, S. Y.; KIM, H. U. Systems strategies for developing industrial microbial strains. *Nat Biotech*, v. 33, n. 10, p. 1061–1072, out. 2015. ISSN 1087-0156.

LEPPäVUORI, J. T.; DOMACH, M. M.; BIEGLER, L. T. Parameter Estimation in Batch Bioreactor Simulation Using Metabolic Models: Sequential Solution with Direct Sensitivities. *Industrial & Engineering Chemistry Research*, v. 50, n. 21, p. 12080–12091, nov. 2011. ISSN 0888-5885, 1520-5045.

LESSIE, T. G.; PHIBBS, P. V. Alternative pathways of carbohydrate utilization in pseudomonads. *Annu. Rev. Microbiol.*, v. 38, p. 359–388, 1984. ISSN 0066-4227.

LEWIS, N. E. et al. Omic data from evolved e. coli are consistent with computed optimal growth from genome-scale models. *Molecular Systems Biology*, v. 6, n. 1, p. 390, 2010.

LLOYD, C. J. et al. COBRAme: A computational framework for genome-scale models of metabolism and gene expression. *PLOS Computational Biology*, v. 14, n. 7, p. e1006302, maio 2018. ISSN 1553-7358. Publisher: Public Library of Science.

LU, H. et al. A consensus S. cerevisiae metabolic model Yeast8 and its ecosystem for comprehensively probing cellular metabolism. *Nature Communications*, v. 10, n. 1, p. 3586, ago. 2019. ISSN 2041-1723. Number: 1 Publisher: Nature Publishing Group.

LYNN, A. R.; SOKATCH, J. R. NOTES Incorporation of Isotope from Specifically Labeled Glucose into Alginates of Pseudomonas aeruginosa and Azotobacter vinelandii. p. 2, 1984.

MACHADO, D.; HERRGåRD, M. Systematic Evaluation of Methods for Integration of Transcriptomic Data into Constraint-Based Models of Metabolism. *PLOS Computational Biology*, v. 10, n. 4, p. e1003580, abr. 2014. ISSN 1553-7358. Publisher: Public Library of Science.

MAHADEVAN, R.; EDWARDS, J. S.; DOYLE, F. J. Dynamic flux balance analysis of diauxic growth in Escherichia coli. *Biophys J*, v. 83, n. 3, p. 1331–1340, set. 2002. ISSN 0006-3495.

MAHADEVAN, R.; SCHILLING, C. H. The effects of alternate optimal solutions in constraint-based genome-scale metabolic models. *Metabolic Engineering*, v. 5, n. 4, p. 264–276, out. 2003. ISSN 1096-7176.

MARANAS, C. D. Optimization Methods in Metabolic Networks. p. 281, 2016.

MARANAS, C. D. et al. SYSTEMS ENGINEERING CHALLENGES AND OPPORTUNITIES IN COMPUTATIONAL BIOLOGY. p. 14, 2003.

MARANAS, C. D.; ZOMORRODI, A. R. *Optimization Methods in Metabolic Networks*. [S.l.]: Wiley, 2016.

MCKINLAY, J. B. et al. Non-growing *Rhodopseudomonas palustris* Increases the Hydrogen Gas Yield from Acetate by Shifting from the Glyoxylate Shunt to the Tricarboxylic Acid Cycle. *Journal of Biological Chemistry*, v. 289, n. 4, p. 1960–1970, jan. 2014. ISSN 0021-9258, 1083-351X.

MCLEAN, K. A. P.; MCAULEY, K. B. Mathematical modelling of chemical processes—obtaining the best model predictions and parameter estimates using identifiability and estimability procedures. *The Canadian Journal of Chemical Engineering*, v. 90, n. 2, p. 351–366, 2012. ISSN 1939-019X.

MEARS, L. et al. A review of control strategies for manipulating the feed rate in fed-batch fermentation processes. *Journal of Biotechnology*, v. 245, p. 34–46, mar. 2017. ISSN 0168-1656.

MENDONCA, C. M.; WILKES, R. A.; ARISTILDE, L. Advancements in 13C isotope tracking of synergistic substrate co-utilization in Pseudomonas species and implications for biotechnology applications. *Current Opinion in Biotechnology*, v. 64, p. 124–133, ago. 2020. ISSN 0958-1669.

MöLLNEY, M. et al. Bidirectional reaction steps in metabolic networks: IV. Optimal design of isotopomer labeling experiments. *Biotechnol. Bioeng.*, v. 66, n. 2, p. 86–103, jan. 1999. ISSN 1097-0290.

NAKAMA, C. S. M.; ROUX, G. A. C. L.; ZAVALA, V. M. Optimal constraint-based regularization for parameter estimation problems. *Computers & Chemical Engineering*, v. 139, p. 106873, ago. 2020. ISSN 0098-1354.

NIELSEN, J. Systems biology of metabolism. *Annual Review of Biochemistry*, v. 86, n. 1, p. 245–275, 2017. PMID: 28301739.

NIELSEN, J.; KEASLING, J. D. Engineering cellular metabolism. v. 164, n. 6, p. 1185–1197, 2016. ISSN 0092-8674, 1097-4172. Publisher: Elsevier.

NIKEL, P. I. et al. *Pseudomonas putida* KT2440 Strain Metabolizes Glucose through a Cycle Formed by Enzymes of the Entner-Doudoroff, Embden-Meyerhof-Parnas, and Pentose Phosphate Pathways. *Journal of Biological Chemistry*, v. 290, n. 43, p. 25920–25932, out. 2015. ISSN 0021-9258, 1083-351X.

NöH, K.; WAHL, A.; WIECHERT, W. Computational tools for isotopically instationary 13c labeling experiments under metabolic steady state conditions. *Metabolic Engineering*, v. 8, n. 6, p. 554–577, nov. 2006. ISSN 1096-7176.

OLAVARRIA, K. et al. Quantifying NAD(P)H production in the upper Entner–Doudoroff pathway from Pseudomonas putida KT2440. *FEBS Open Bio*, v. 5, n. 1, p. 908–915, jan. 2015. ISSN 2211-5463.

OLIVEIRA, R. D. et al. Nonlinear Predictive Control of a Bioreactor by Surrogate Model Approximation of Flux Balance Analysis. *Industrial & Engineering Chemistry Research*, out. 2021. ISSN 0888-5885. Publisher: American Chemical Society.

OLIVEIRA, R. D. et al. Identifiability of metabolic flux ratios on carbon labeling experiments. In: *Computer Aided Chemical Engineering*. [S.l.]: Elsevier, 2021. v. 50, p. 1983–1989. ISBN 978-0-323-88506-5.

OLIVEIRA, R. D. et al. Glucose metabolism in Pseudomonas aeruginosa is cyclic when producing Polyhydroxyalkanoates and Rhamnolipids. *Journal of Biotechnology*, v. 342, p. 54–63, dez. 2021. ISSN 01681656.

OLIVEIRA, R. D. et al. Parameter estimation in dynamic metabolic models applying a surrogate approximation. 2022.

OLIVEIRA, R. D. d. *Desenvolvimento de uma ferramenta computacional para a análise de fluxos metabólicos empregando carbono marcado.* Tese (Mestrado em Engenharia Química) — Universidade de São Paulo, São Paulo, jan. 2018.

OPPERMAN, M. J.; SHACHAR-HILL, Y. Metabolic flux analyses of Pseudomonas aeruginosa cystic fibrosis isolates. *Metabolic Engineering*, v. 38, n. Supplement C, p. 251–263, nov. 2016. ISSN 1096-7176.

ORTH, J. D. et al. A comprehensive genome-scale reconstruction of Escherichia coli metabolism—2011. *Molecular Systems Biology*, v. 7, p. 535, out. 2011. ISSN 1744-4292.

ORTH, J. D.; THIELE, I.; PALSSON, B. What is flux balance analysis? *Nature Biotechnology*, v. 28, n. 3, p. 245–248, mar. 2010. ISSN 1087-0156, 1546-1696.

PALSSON, B. *Systems Biology: Constraint-based Reconstruction and Analysis*. [S.l.]: Cambridge University Press, 2015.

PAULSON, J. A.; MARTIN-CASAS, M.; MESBAH, A. Fast uncertainty quantification for dynamic flux balance analysis using non-smooth polynomial chaos expansions. *PLOS Computational Biology*, v. 15, n. 8, p. e1007308, ago. 2019. ISSN 1553-7358. Publisher: Public Library of Science.

PEIXOTO, R. M. Bioprospecção de microrganismos do gênero Pseudomonas produtores de biossurfactantes. p. 98, 2008.

PFLüGER, D.; PEHERSTORFER, B.; BUNGARTZ, H.-J. Spatially adaptive sparse grids for high-dimensional data-driven problems. v. 26, n. 5, p. 508–522, 2010. ISSN 0885-064X.

PLOCH, T. et al. Simulation of differential-algebraic equation systems with optimization criteria embedded in Modelica. *Computers & Chemical Engineering*, v. 140, p. 106920, set. 2020. ISSN 0098-1354.

PLUS, A. Aspen plus user guide. *Aspen Technology Limited, Cambridge, Massachusetts, United States*, 2003.

PORTAIS, J.-C.; DELORT, A.-M. Carbohydrate cycling in micro-organisms: what can ^{13}C-NMR tell us? *FEMS Microbiology Reviews*, v. 26, n. 4, p. 375–402, nov. 2002. ISSN 1574-6976.

QUEIPO, N. V. et al. Surrogate-based analysis and optimization. *Progress in Aerospace Sciences*, v. 41, n. 1, p. 1–28, 2005. ISSN 0376-0421.

RACKAUCKAS, C.; NIE, Q. Differentialequations.jl – a performant and feature-rich ecosystem for solving differential equations in julia. *Journal of Open Research Software*, v. 5, 05 2017.

RAGHUNATHAN, A. U. et al. Parameter estimation in metabolic flux balance models for batch fermentation—Formulation & Solution using Differential Variational Inequalities (DVIs). *Annals of Operations Research*, v. 148, n. 1, p. 251–270, nov. 2006. ISSN 0254-5330, 1572-9338.

RAGHUNATHAN, A. U.; PéREZ-CORREA, J. R.; BIEGER, L. T. Data reconciliation and parameter estimation in flux-balance analysis. *Biotechnology and Bioengineering*, v. 84, n. 6, 2003. ISSN 1097-0290.

RAJ, K.; VENAYAK, N.; MAHADEVAN, R. Novel two-stage processes for optimal chemical production in microbes. *Metabolic Engineering*, v. 62, p. 186–197, nov. 2020. ISSN 10967176.

RAMKRISHNA, D.; SONG, H.-S. *Cybernetic Modeling for Bioreaction Engineering*. 1. ed. [S.l.]: Cambridge University Press, 2018. ISBN 978-0-511-73196-9 978-1-107-00052-0.

RAMSAY, B. A. et al. Production of poly-(beta-hydroxybutyric-co-beta-hydroxyvaleric) acids. *Applied and Environmental Microbiology*, v. 56, n. 7, p. 2093–2098, 1990. ISSN 0099-2240.

RANDHAWA, K. K. S.; RAHMAN, P. K. S. M. Rhamnolipid biosurfactants—past, present, and future scenario of global market. *Front Microbiol*, v. 5, set. 2014. ISSN 1664-302X.

RAWLINGS, J.; MAYNE, D.; DIEHL, M. *Model Predictive Control: Theory, Computation, and Design*. [S.l.]: Nob Hill Publishing, 2017. ISBN 9780975937730.

RAZA, Z. A.; ABID, S.; BANAT, I. M. Polyhydroxyalkanoates: Characteristics, production, recent developments and applications. *International Biodeterioration & Biodegradation*, v. 126, p. 45–56, jan. 2018. ISSN 0964-8305.

REED, J. L. et al. An expanded genome-scale model of Escherichia coli K-12 (iJR904 GSM/GPR). *Genome Biology*, v. 4, n. 9, p. R54, ago. 2003. ISSN 1474-760X.

REHM, B. H. A.; MITSKY, T. A.; STEINBüCHEL, A. Role of Fatty Acid De Novo Biosynthesis in Polyhydroxyalkanoic Acid (PHA) and Rhamnolipid Synthesis by Pseudomonads: Establishment of the Transacylase (PhaG)-Mediated Pathway for PHA Biosynthesis inEscherichia coli. *Appl. Environ. Microbiol.*, v. 67, n. 7, p. 3102–3109, jul. 2001. ISSN 0099-2240, 1098-5336.

RIASCOS, C. A. M. et al. Metabolic pathways analysis in PHAs production by Pseudomonas with 13C-labeling experiments. *Computer Aided Chemical Engineering*, v. 32, p. 121–126, jan. 2013. ISSN 1570-7946.

RIIS, V.; MAI, W. Gas chromatographic determination of poly--hydroxybutyric acid in microbial biomass after hydrochloric acid propanolysis. *Journal of Chromatography A*, v. 445, p. 285 – 289, 1988. ISSN 0021-9673.

SAA, P. A.; NIELSEN, L. K. Formulation, construction and analysis of kinetic models of metabolism: A review of modelling frameworks. *Biotechnology Advances*, v. 35, n. 8, p. 981–1003, dez. 2017. ISSN 0734-9750.

SAUER, U. Metabolic networks in motion: 13C-based flux analysis. *Molecular Systems Biology*, v. 2, nov. 2006. ISSN 1744-4292.

SCHUSTER, S.; HLGETAG, C. ON ELEMENTARY FLUX MODES IN BIOCHEMICAL REACTION SYSTEMS AT STEADY STATE. *Journal of Biological Systems*, v. 2, n. 2, p. 165–182, mar. 1994.

SCOTT, F. et al. Simulation and optimization of dynamic flux balance analysis models using an interior point method reformulation. *Computers Chemical Engineering*, v. 119, p. 152–170, 2018. ISSN 0098-1354.

SHIN, S.; VENTURELLI, O. S.; ZAVALA, V. M. Scalable nonlinear programming framework for parameter estimation in dynamic biological system models. *PLOS Computational Biology*, v. 15, n. 3, p. e1006828, mar. 2019. ISSN 1553-7358. Publisher: Public Library of Science.

SIGMA-ALDRICH. *D-Glucose-1*. 2021. Disponível em: <https://www.sigmaaldrich.com/catalog/product/aldrich-/297046?lang=pt&region=BR&gclid=CjwKCAjwxev3BRBBEiwAiB_PWMJdOeTkCV0wKTxstfHsEkJ6PbfX

SINGH, D.; LERCHER, M. J. Network reduction methods for genome-scale metabolic models. v. 77, n. 3, p. 481–488, 2021. ISSN 1420-9071.

SOMMEREGGER, W. et al. Quality by control: Towards model predictive control of mammalian cell culture bioprocesses. *Biotechnology journal*, Wiley Online Library, v. 12, n. 7, p. 1600546, 2017.

SRINIVASAN, S.; CLUETT, W. R.; MAHADEVAN, R. Constructing kinetic models of metabolism at genome-scales: A review. *Biotechnology Journal*, v. 10, n. 9, p. 1345–1359, 2015. ISSN 1860-7314.

SROUR, O.; YOUNG, J. D.; ELDAR, Y. C. Fluxomers: a new approach for 13 C metabolic flux analysis. *BMC systems biology*, v. 5, n. 1, p. 1, 2011.

STEPHANOPOULOS, G. Metabolic engineering. *Current Opinion in Biotechnology*, v. 5, n. 2, p. 196–200, 1994.

STRAUS, J.; SKOGESTAD, S. A new termination criterion for sampling for surrogate model generation using partial least squares regression. v. 121, p. 75–85, 2019. ISSN 0098-1354.

SáNCHEZ-PASCUALA, A. et al. Functional implementation of a linear glycolysis for sugar catabolism in Pseudomonas putida. *Metabolic Engineering*, abr. 2019. ISSN 1096-7176.

THEORELL, A. et al. To be certain about the uncertainty: Bayesian statistics for 13C metabolic flux analysis. *Biotechnology and Bioengineering*, v. 114, n. 11, 2017. ISSN 1097-0290.

THIELE, I.; PALSSON, B. A protocol for generating a high-quality genome-scale metabolic reconstruction. *Nat Protoc*, v. 5, n. 1, p. 93–121, 2010. ISSN 1754-2189.

TSIANTIS, N.; BALSA-CANTO, E.; BANGA, J. R. Optimality and identification of dynamic models in systems biology: an inverse optimal control framework. *Bioinformatics*, v. 34, n. 14, p. 2433–2440, jul. 2018. ISSN 1367-4803.

VAJDA, S. et al. Qualitative and quantitative identifiability analysis of nonlinear chemical kinetic models. *Chemical Engineering Communications*, v. 83, n. 1, p. 191–219, set. 1989. ISSN 0098-6445, 1563-5201.

VASILAKOU, E. et al. Current state and challenges for dynamic metabolic modeling. v. 33, p. 97–104, 2016. ISSN 1369-5274.

WÄCHTER, A.; BIEGLER, L. T. On the implementation of an interior-point filter line-search algorithm for large-scale nonlinear programming. *Mathematical programming*, Springer, v. 106, n. 1, p. 25–57, 2006.

WALDHERR, S. State estimation in constraint based models of metabolic-genetic networks. In: *2016 American Control Conference (ACC)*. [S.l.: s.n.], 2016. p. 6683–6688. ISSN: 2378-5861.

WIECHERT, W. Algebraic Methods for the Analysis of Redundancy and Identifiability in Metabolic 13C-Labehg Systems. *Bioinformatics: From Nucleic Acids and Proteins to Cell Metabolism*, v. 18, p. 85, 1995.

WIECHERT, W. 13c Metabolic Flux Analysis. *Metabolic Engineering*, mar. 2001.

WIECHERT, W. The Thermodynamic Meaning of Metabolic Exchange Fluxes. *Biophysical Journal*, v. 93, n. 6, p. 2255–2264, set. 2007. ISSN 0006-3495.

WIECHERT, W.; GRAAF, A. A. de. Bidirectional reaction steps in metabolic networks: I. Modeling and simulation of carbon isotope labeling experiments. *Biotechnol. Bioeng.*, v. 55, n. 1, p. 101–117, jul. 1997. ISSN 1097-0290.

WIECHERT, W.; NöH, K. Isotopically non-stationary metabolic flux analysis: complex yet highly informative. v. 24, n. 6, p. 979–986, 2013. ISSN 0958-1669.

WIECHERT, W. et al. Bidirectional reaction steps in metabolic networks: II. Flux estimation and statistical analysis. *Biotechnol. Bioeng.*, v. 55, n. 1, p. 118–135, jul. 1997. ISSN 1097-0290.

WINDEN, W. A. van; HEIJNEN, J. J.; VERHEIJEN, P. J. Cumulative bondomers: A new concept in flux analysis from 2d [13c,1h] COSY NMR data. *Biotechnology and Bioengineering*, v. 80, n. 7, p. 731–745, dez. 2002. ISSN 0006-3592, 1097-0290. Disponível em: <9>.

WINDEN, W. A. van et al. A priori analysis of metabolic flux identifiability from 13C-labeling data. *Biotechnology and Bioengineering*, v. 74, n. 6, p. 505–516, set. 2001. ISSN 1097-0290.

WINDEN, W. van; VERHEIJEN, P.; HEIJNEN, S. Possible pitfalls of flux calculations based on 13c-labeling. v. 3, n. 2, p. 151–162, 2001. ISSN 1096-7176.

WURZEL, M.; GRAAF, A. A. de. Bidirectional Reaction Steps in Metabolic Networks: III. Explicit Solution and Analysis of Isotopomer Labeling Systems. *BIOTECHNOLOGY AND BIOENGINEERING*, v. 66, n. 2, 1999.

ZETTERHOLM, J. et al. Economic evaluation of large-scale biorefinery deployment: A framework integrating dynamic biomass market and techno-economic models. v. 12, n. 17, p. 7126, 2020. ISSN 2071-1050. Number: 17 Publisher: Multidisciplinary Digital Publishing Institute.

ZHAO, X. et al. Dynamic flux balance analysis with nonlinear objective function. *Journal of Mathematical Biology*, v. 75, n. 6, p. 1487–1515, dez. 2017. ISSN 1432-1416.

ZHENG, H.; RICARDEZ-SANDOVAL, L.; BUDMAN, H. Robust estimation and economic predictive control for dynamic metabolic flux systems under probabilistic uncertainty. v. 140, p. 106918, 2020. ISSN 0098-1354.

ZHU, K.; ROCK, C. O. RhlA Converts -Hydroxyacyl-Acyl Carrier Protein Intermediates in Fatty Acid Synthesis to the -Hydroxydecanoyl--Hydroxydecanoate Component of Rhamnolipids in Pseudomonas aeruginosa. *J Bacteriol*, v. 190, n. 9, p. 3147–3154, maio 2008. ISSN 0021-9193.

ZHUANG, K. et al. Genome-scale dynamic modeling of the competition between Rhodoferax and Geobacter in anoxic subsurface environments. *The ISME journal*, v. 5, n. 2, p. 305–316, fev. 2011. ISSN 1751-7362.

ZUPKE, C.; STEPHANOPOULOS, G. Modeling of isotope distributions and intracellular fluxes in metabolic networks using atom mapping matrices. *Biotechnology Progress*, v. 10, n. 5, p. 489–498, 1994.

ÖNER, M.; STOCKS, S. M.; SIN, G. Comprehensive sensitivity analysis and process risk assessment of large scale pharmaceutical crystallization processes. *Computers Chemical Engineering*, v. 135, p. 106746, 2020. ISSN 0098-1354.

Appendices

Table 13 – Atoms transitions in metabolic network model of of *Pseudomonas aeruginosa LFM634*.

Reaction	Stoichiometry	Atom carbon transition
1	Glucose → G6P	abcdef → abcdef
2	Glucose → Gluconate	abcdef → abcdef
3	Gluconate → 6PG	abcdef → abcdef
4	G6P → 6PG	abcdef → abcdef
5	G6P → Rhamnose	abcdef → abcdef
6	6PG → G3P + PYR	abcdef → def + abc
7	G3P + G3P → F6P	abc + def → cbadef
8	F6P ⇔ G6P	abcdef ⇔ abcdef
9	P5P + E4P ⇔ G3P + F6P	abcde + fghi ⇔ cde + abfghi
10	G3P + S7P ⇔ E4P + F6P	abc + defghij ⇔ ghij + defabc
11	P5P + P5P ⇔ S7P + G3P	abcde + fghij ⇔ abfghij + cde
12	G3P → PYR	abc→abc
13	PYR → AcCoA	abc→bc
14	AcCoA + AcCoA→ 3HA/3HAA	ab + cd→ abc

Table 14 – GC-MS analysis of monomers 3HD in PHA and 3HAA in fragments m/z=89 and m/z=131 for medium C/N=45. Medium with a mixture of 55% (w/w) [6-13C]glucose and 45% of natural glucose.

	PHA		3HAA	
Mass	m/z=89		m/z=89	
isotopomer	mean	s.d.	mean	s.d.
M+0	0.8397	0.0023	0.8387	0.0031
M+1	0.1457	0.0020	0.1463	0.0032
M+2	0.0127	0.0003	0.0132	0.0007
M+3	0.0019	0.0001	0.0017	0.0008

Table 15 – GC-MS analysis of monomers 3HD in PHA and 3HAA in fragment m/z=89 for medium C/N=45. Medium with a mixture of 80% (w/w) [U-^{13}C]glucose and 20% of natural glucose. standart deviation (s.d.)

	PHA		3HAA	
Mass	C/N=45		C/N=45	
isotopomer	mean	s.d.	mean	s.d.
M+0	0.6294	0.0090	0.6270	0.0054
M+1	0.2080	0.0028	0.2093	0.0032
M+2	0.1339	0.0054	0.1255	0.0026
M+3	0.0288	0.0014	0.0281	0.0010

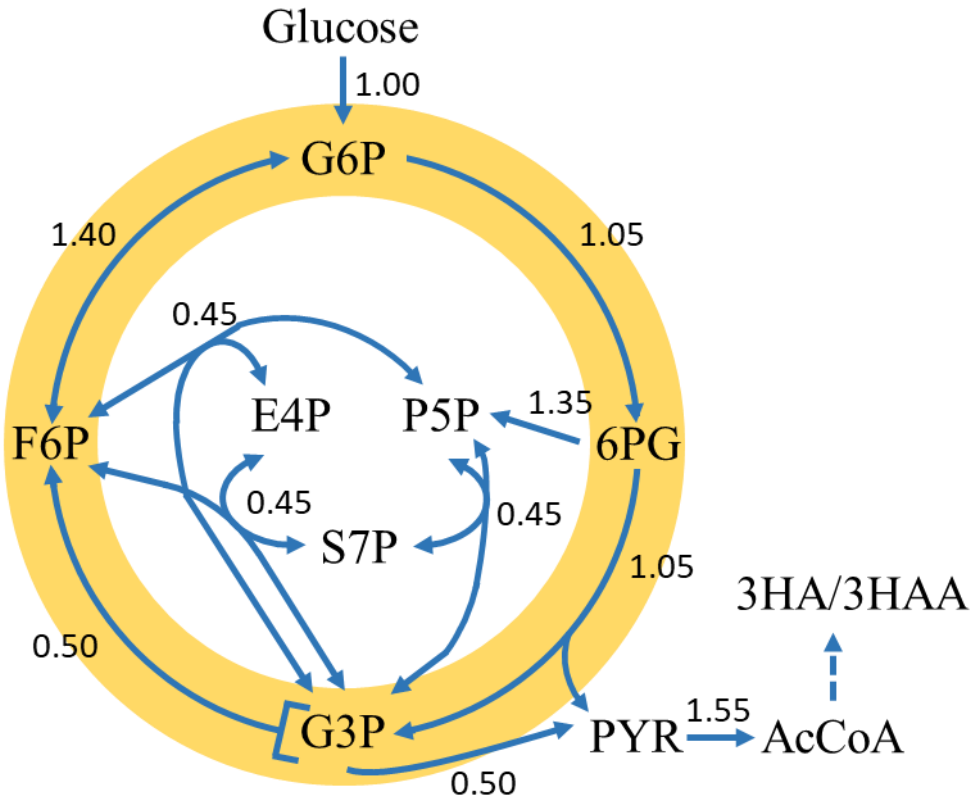


Figure 32 – *a priori* flux distribution on *Pseudomonas spp.* central glucose metabolism metabolic network.

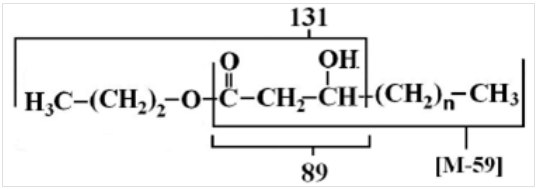


Figure 33 – Important mass fragments in 3HAs and 3HAA monomers.

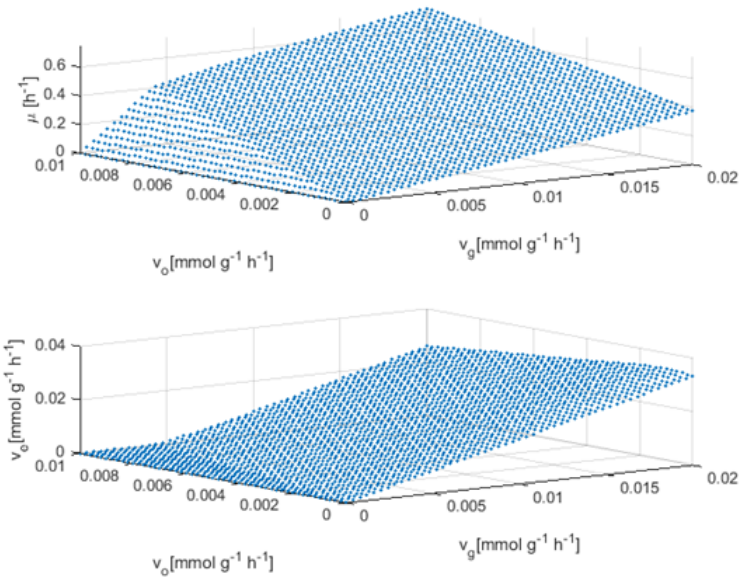


Figure 34 – Profiles of the FBA solutions for different values of the uptake rates of glucose and of oxygen. We solved the FBA problem for every value of the independent variables in an equidistant 50 by 50 grid. Here, only one zone of operation was used.

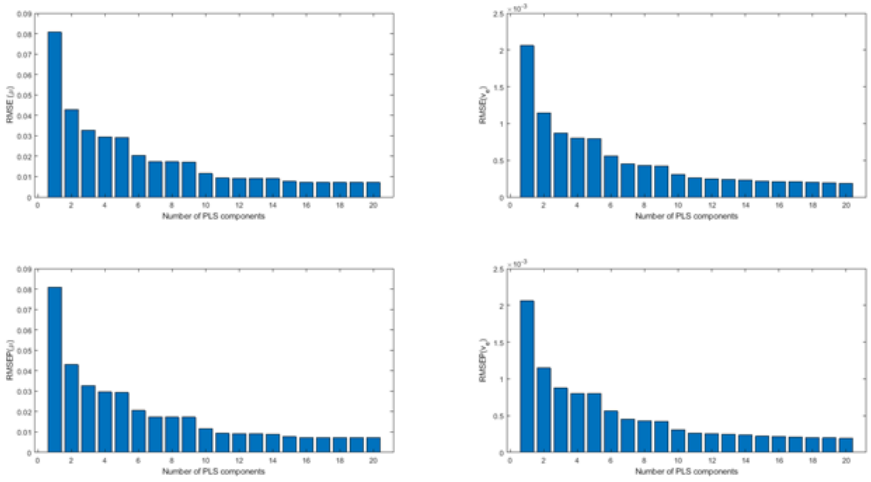


Figure 35 – Relative RMSE and RMSEP for growth rate model and the exchange ethanol rate as a function of the number of PLS components for a surrogate model with a single zone (Figure 34).

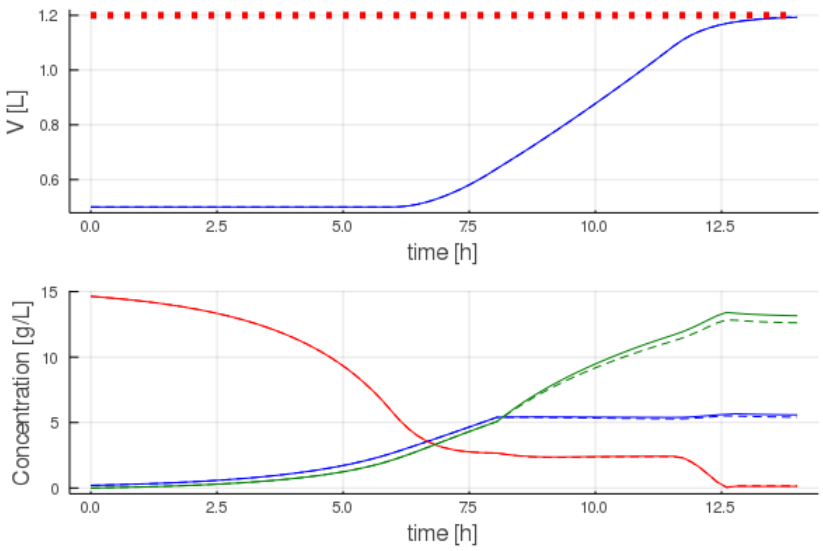


Figure 36 – Profiles obtained using the surrogate model with a single zone (–) and three zones (-) in the controller. The standard dFBA was used as the bioreactor model. The concentration profiles are glucose (red), biomass (blue) and ethanol (green).

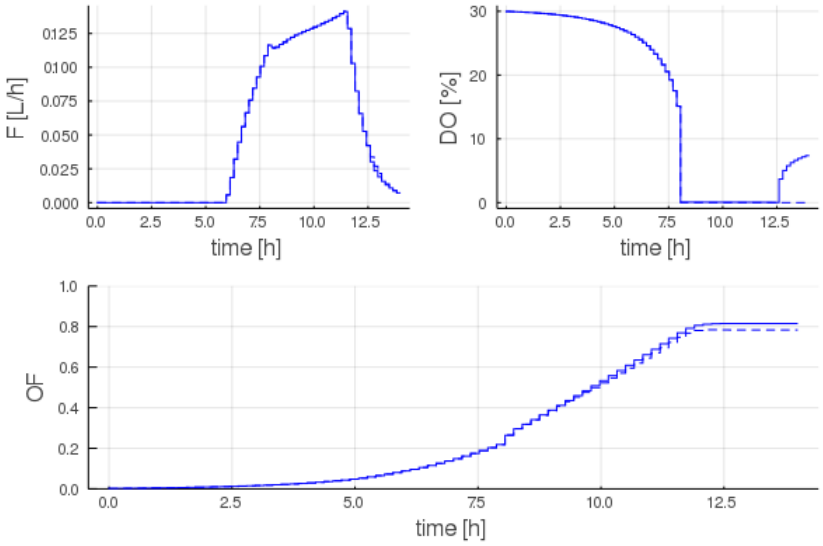


Figure 37 – Control variables and objective function (OF) profiles obtained using the surrogate model with a single zone (–) and three zones (-) in the controller. The standard dFBA was used as the bioreactor model.

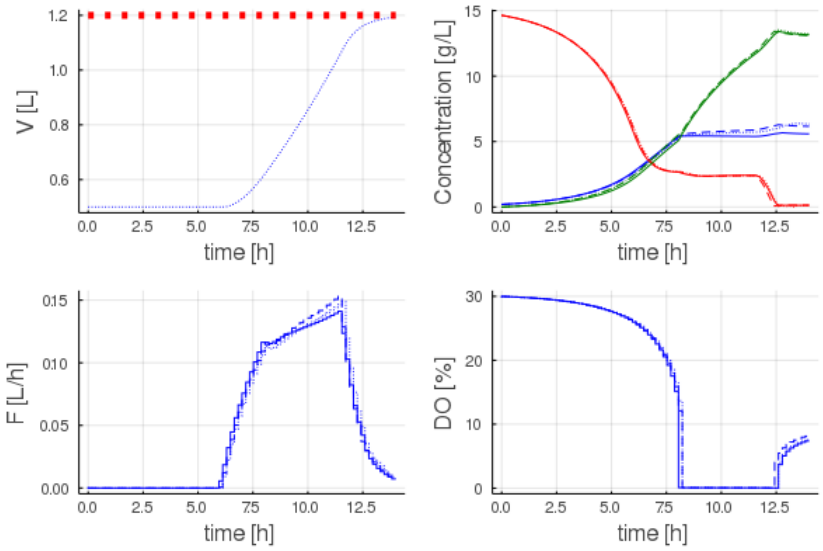


Figure 38 – Comparison between different GSM for the plant-model: Yeast 8.3 (-), iND750 (–) and iMM904 (···). The simulations were performed on open-loop operation. The concentration profiles are glucose (red), biomass (blue) and ethanol (green).