PONTIFICIA UNIVERSIDAD CATOLICA DE CHILE
ESCUELA DE INGENIERIA

DYNAMIC GENOME-SCALE METABOLIC MODELING OF THE YEAST *PICHIA PASTORIS*

FRANCISCO JAVIER SAITUA

Thesis submitted to the Office of Research and Graduate Studies in partial fulfillment of the requirements for the Degree of Master of Science in Engineering

Advisor:
EDUARDO AGOSIN TRUMPER

Santiago of Chile, August, 2016

PONTIFICIA UNIVERSIDAD CATOLICA DE CHILE
ESCUELA DE INGENIERIA

DYNAMIC GENOME-SCALE METABOLIC MODELING OF THE YEAST *PICHIA PASTORIS*

FRANCISCO JAVIER SAITUA PÉREZ

Members of the Committee:

EDUARDO AGOSIN TRUMPER

JOSÉ RICARDO PÉREZ CORREA

MARÍA RODRÍGUEZ-FERNÁNDEZ

ALEJANDRO MAASS

JOSE MIGUEL CEMBRANO PERASSO

Thesis submitted to the Office of Research and Graduate Studies in partial fulfillment of the requirements for the Degree of Master of Science in Engineering

Santiago de Chile, August, 2016

To my parents,

example of love, compassion

and intelligence

at the service of others.

ACKNOWLEDGEMENTS

I want to express my sincere gratitude to my supervisor, Prof. Eduardo Agosín, for his support and guidance throughout this process. The entertaining discussions and challenging tasks that emerged from them were crucial for the main findings of this study. I would like to extend this to Professor Ricardo Pérez, who provided mathematical support from the beginning of my work, giving me ideas to overcome several numerical issues.

I would also like to thank the rest of the members of the committee, Professors José Miguel Cembrano, María Rodríguez-Fernández and Alejandro Maass, for their time to correct the final manuscript and the thoughtful comments made on it.

Furthermore, I want to acknowledge the Graduate Program of the School of Engineering from Pontificia Universidad Católica de Chile for their help in the application for the CONICYT scholarship and for the financial support provided for my assistance to an important congress of my field.

This thesis would not have been possible without the help of my friends from the Biotechnology Lab directed by Professor Agosín. Each one of you helped me in some way to fulfill this task, either with your questions during seminars, help during the experiments or nurturing discussions during launch. I would like to give special thanks to Paulina Torres, my friend and co-worker in the *P. pastoris* group, for her help in the experiments, analysis of results and thoughtful comments on the final manuscript.

It is also important for me to acknowledge the personnel of the Department of Chemical and Bioprocess Engineering, who also played a relevant role in the fulfillment of this thesis. I deeply appreciate the technical support provided during the experiments, the clean working environment I found every time I came to the Lab and the help given to rapidly complete administrative tasks.

Finally, I would like to thank my family for their love, unconditional support and for giving me the opportunity to get properly educated, I do not take that for granted.

GENERAL INDEX

FIGURE INDEX

TABLE INDEX

ABSTRACT

The yeast *Pichia pastoris* shows physiological advantages for the production of recombinant proteins compared to other commonly used cell factories. This microorganism is mostly grown in dynamic conditions, where the cell's environment is continuously changing and many variables influence process productivity. In this context, a model capable of explaining and predicting cell behavior for the rational design of bioprocesses is highly desirable.

In this work, we developed a dynamic genome-scale metabolic model for glucose-limited, aerobic cultivations of *Pichia pastoris*. Starting from an initial structure for batch and fed-batch configurations, we performed pre/post regression diagnostics to determine identifiability, significance and sensitivity problems between model parameters. Once identified, the non-relevant parameters were iteratively fixed until an *a priori* robust modeling structure was found for both types of cultivation. Next, the robustness of these structures was confirmed by calibrating new datasets, where no parametric problems appeared. Finally, the model was validated for the prediction of batch and fed-batch dynamics in the studied conditions.

The platform was also used to unravel genetic and process engineering strategies to improve the production recombinant Human Serum Albumin (HSA). The simulation of single knock-outs indicated that the deviation of carbon towards cysteine and tryptophan formation could theoretically improve HSA production. In particular, the deletion of methylene tetrahydrofolate dehydrogenase could potentially increase HSA volumetric productivity by 630%. Also, given specific bioprocess limitations and strain characteristics, the model indicated that a decreasing growth rate in the feed phase of a fed-batch cultivation may improve the volumetric productivity of this protein by 24%.

We formulated a dynamic genome scale metabolic model of *Pichia pastoris* that yields realistic metabolic flux distributions throughout dynamic cultivations. It can be used to calibrate several experimental data and to rationally propose metabolic and process engineering strategies to improve the performance of a cultivation setup.

RESUMEN

Pichia pastoris posee ventajas fisiológicas para la producción de proteínas recombinantes en comparación a las factorías celulares convencionales. Esta levadura es comúnmente cultivada en condiciones dinámicas, donde el ambiente extracelular cambia constantemente y numerosas variables inciden en la productividad del proceso. En este contexto, un modelo capaz de explicar y predecir el comportamiento celular para el diseño racional de bioprocesos es altamente deseable.

En este trabajo, se ensambló un modelo dinámico a escala genómica de *Pichia pastoris* para cultivos aeróbicos *batch* y *fed-batch* limitados en glucosa. El modelo fue calibrado con datos de fermentaciones, luego de lo cual se realizaron diagnósticos de pre/post regresión para determinar problemas de sensibilidad, significancia e identificabilidad entre sus parámetros. Una vez identificados, los parámetros irrelevantes fueron fijados iterativamente hasta encontrar una estructura de modelación sin problemas paramétricos *a priori*. La robustez de estas estructuras fue comprobada mediante la calibración de nuevos datos experimentales, donde no aparecieron los problemas antes mencionados. Finalmente, el modelo fue validado para la predicción de dinámicas *batch* y *fed-batch* en condiciones similares a las estudiadas.

Luego de la validación, el modelo fue utilizado para revelar estrategias de ingeniería genética y de procesos para optimizar la producción de Albúmina Sérica Humana (HSA) recombinante. La simulación de *knock-outs* indicó que el desvío del carbono hacia la formación de cisteína y triptófano podría mejorar la producción de HSA. En particular, la deleción de la enzima Metilen-tetrahidrofolato deshidrogenasa podría aumentar la productividad volumétrica de la formación de HSA en un 630%. Además, el modelo indicó que es posible mejorar en un 24% la productividad de la formación de HSA mediante una política de tasa de crecimiento decreciente en la fase de alimentación de un cultivo *fed-batch*.

En conclusión, se formuló un modelo dinámico a escala genómica de *Pichia pastoris* capaz de entregar distribuciones de flujo realistas durante cultivos dinámicos y proponer,

de manera racional, estrategias de ingeniería metabólica y de procesos para mejorar el desempeño de un biorreactor.

Keywords: dFBA, *Pichia pastoris*, Pre/post regression diagnostics, Sensitivity, Identifiability, Significance, Genome-scale metabolic modeling, fed-batch, MOMA, Bioprocess optimization, Reparametrization.

1. INTRODUCTION

Recombinant protein production is a multibillion-dollar business mainly comprised by biopharmaceuticals (i.e. recombinant biologic drugs) and industrial enzymes (BCC Research, 2014; Markets and Markets, 2015; Walsh, 2014). These compounds are commonly synthesized in *Escherichia coli*, *Saccharomyces cerevisiae* and Chinese Hamster Ovary cells (CHO) (Ferrer-Miralles, Domingo-Espín, Corchero, Vázquez, & Villaverde, 2009; Maccani et al., 2014; Overton, 2014; Walsh, 2014); however, there is a strong pressure to find cost-effective alternatives to overcome technical and economic disadvantages of the aforementioned cell factories, especially in downstream processing (Corchero et al., 2013).

Among the unconventional cell factories used for recombinant protein production, the methylotrophic yeast *Pichia pastoris* (syn. *Komagataella phaffii*) has driven special attention thanks to its convenient physiology and easy handling (Daly & Hearn, 2005). First of all, commercially available strong promoters (inducible and constitutive) allow the controlled expression of heterologous proteins easily (Daly & Hearn, 2005). Unlike *E. coli*, *P. pastoris* naturally performs post-translational modifications such as disulfide bond formation, proteolytic processing and glycosylation (Cereghino & Cregg, 2000; Ferrer-Miralles et al., 2009). This feature enables the protein being produced to achieve a correct tertiary structure, which is essential for its functionality (Ciofalo, Barton, Kreps, Coats, & Shanahan, 2006; Corchero et al., 2013; Masuda, Ide, Ohta, & Kitabatake, 2010). In contrast to *S. cerevisiae*, *P. pastoris* exhibits a Crabtree-negative phenotype, showing a reduced formation of undesirable products, like ethanol, in glucose-limited conditions (Çalık et al., 2015; Mattanovich et al., 2009). It also shows, when compared to other yeasts, a lower basal secretion of proteins, which makes downstream processing easier (Delic et al., 2013; Mattanovich et al., 2009). Finally, it can be efficiently grown up to high cell densities using fed-batch cultivations (Daly & Hearn, 2005), achieving high titers and productivities. For these desirable features, *P. pastoris* has been widely used for the expression of recombinant proteins, reaching grams per liter concentrations in several cases (Cereghino & Cregg, 2000; Čiplys et al., 2015; Hasslacher et al., 1997; Heyland, Fu, Blank, & Schmid, 2010; Y. Wang et al., 2001). Most remarkably, and as a proof of its

technical feasibility and adequacy, two recombinant proteins produced in this cell factory have already been approved by the FDA for medical purposes (Ciofalo et al., 2006; Thompson, 2010).

Despite its growing acceptance and actual successful applications, recombinant protein production in *P. pastoris* can be undermined by several cellular processes, with protein folding and secretion being the most recurrent bottlenecks (Delic et al., 2013; Delic, Göngrich, Mattanovich, & Gasser, 2014; Gasser et al., 2013). In addition, limitations may also be caused by the codon usage of the recombinant protein (J.-R. Wang et al., 2015), promoter selection (Prielhofer et al., 2013), carbon and oxygen availability in the culture (Baumann et al., 2008; Heyland, Fu, Blank, & Schmid, 2011) and fed-batch operational parameters (Maurer, Kühleitner, Gasser, & Mattanovich, 2006), seriously hampering protein yield, productivity and the economic feasibility of the process.

Industrially, *Pichia pastoris* is commonly grown in fed-batch cultures (Looser et al., 2015). This configuration allows to reach high titers of a product of interest with limited formation of undesirable compounds in a controlled fashion (Villadsen, Nielsen, & Lidén, 2011). During this type of cultivation, the extracellular medium changes constantly and the cells adapt continuously to the varying concentration of species. In this context, it is important to understand the metabolic impact that process conditions generate in the cell to improve the strain's performance (Graf, Dragosits, Gasser, & Mattanovich, 2009). This is a complex task since process variables often require significant amounts of time - and money - to characterize and fine-tune (Looser et al., 2015). Therefore, it would desirable to have a computer platform for *P. pastoris* that allows the simulation of its growth under different experimental setups in order to elaborate rational hypotheses to improve the production of a compound of interest.

Systems biology offers a quantitative and comprehensive approach to address this task (Kitano, 2002). In particular, dynamic Flux Balance Analysis (dFBA) (Mahadevan, Edwards, & Doyle, 2002; Varma & Palsson, 1994) is a modeling framework that allows the simulation of non-stationary (batch or fed-batch) cultures of a target microorganism, accounting for its entire metabolism using Genome-Scale Metabolic Models (GSM).

GSMs are computable structures that represent the entire metabolism of a cell or microbial community (Palsson, 2015; Thiele & Palsson, 2010). Their applications include understanding cellular behavior under different environmental conditions, serving as a scaffold to map over them omics data and determining favorable culture conditions and genetic modifications for a particular metabolic engineering objective *in silico* (Asadollahi et al., 2009; Park, Lee, Kim, & Lee, 2007). There are five published GSMs of *Pichia pastoris* (Caspeta, Shoaie, Agren, Nookaew, & Nielsen, 2012; B. K. S. Chung et al., 2010; Irani, Kerkhoven, Shojaosadati, & Nielsen, 2015; Sohn et al., 2010; Tomàs-Gamisans, Ferrer, & Albiol, 2016), all designed to help in the strain optimization process with a special emphasis on recombinant protein production. However, their application for this purpose has been limited (B. K.-S. Chung, Lakshmanan, Klement, Ching, & Lee, 2013; Nocon et al., 2014), since they have mostly been employed to study stationary conditions and little attention has been given to the flux distributions that can be derived from them.

2. HYPOTHESIS

The development of a dynamic genome-scale metabolic model of *Pichia pastoris* will enable the determination of metabolic flux distributions during batch and fed-batch cultivations, which can then be used to unravel metabolic and process engineering strategies for the overproduction of a compound of interest.

3. OBJECTIVES

3.1. General Objective

To assemble a dynamic, genome scale metabolic model of *Pichia pastoris* capable of fitting experimental data without parametric problems and that can be used to generate metabolic and process engineering strategies to improve bioreactor performance.

3.2. Specific Objectives:

3.2.1. To assemble a dynamic genome-scale metabolic model for aerobic, glucose-limited dynamic cultivations of *P. pastoris*.

3.2.2. To be able to calibrate new data with the model with as few sensitivity, identifiability and significance problems as possible.

3.2.3. To apply the model to (i) obtain metabolic flux distributions during dynamic cultivations and (ii) to unravel genetic and process engineering strategies to improve the production of a recombinant protein.

4. MATERIALS AND METHODS

To assemble the dynamic modeling framework (), we started by selecting one of the available genome-scale metabolic models and employed it at the core of the model. Once built, the model was calibrated using experimental data from several batch and fed-batch cultivations. Then, its structure was evaluated in order to determine the presence of parametric problems: lack of significance, low sensitivity or non-identifiability (correlation). If these problems are not properly assessed, they can mask the real value of the parameters, which are inputs of the GSM to obtain flux distributions. Once identified, the aforementioned problems were eliminated by iteratively fixing the non-relevant ones, leaving subsets with no issues *a priori*. These subsets were used to recalibrate the available data and the one that presented the best fitting capability with the fewest parametric problems was chosen to be validated as a robust modeling structure. To do this, we first demonstrated that the chosen model structure yielded no (of just a few) significance, sensitivity or identifiability problems when calibrating new data. Complementary to this analysis, we determined if the model could predict accurately bioreactor dynamics in both batch and fed-batch configurations.

Finally, we manually curated the chosen metabolic model and performed simulations to demonstrate the potential uses of the model:

i) Analysis of the metabolic flux distribution during different stages of a dynamic cultivation
ii) Discovery of knock-out targets for the overproduction of recombinant proteins
iii) Evaluation of different feeding policies *in silico* to improve recombinant protein production considering specific information about the strain and process setup.

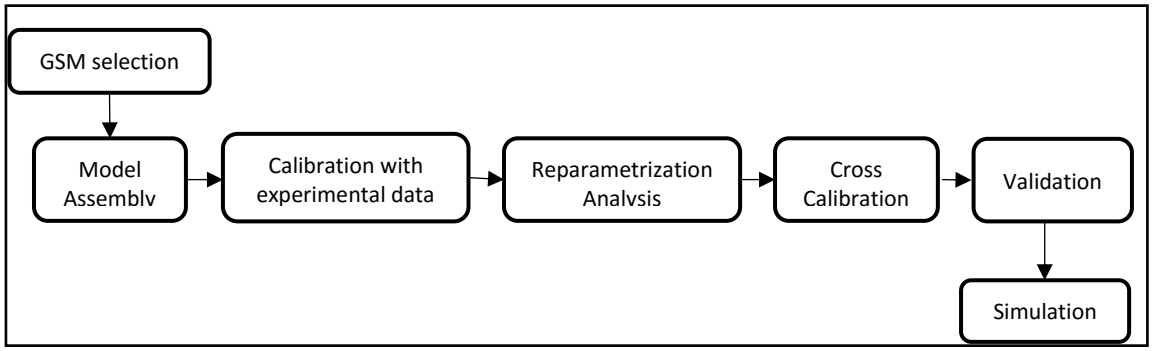


Figure 1 – Methodology workflow.

4.1. Genome-scale model selection from literature

Usability and similarity to experimental chemostat data were used as criteria to select the most appropriate genome-scale model for building the dynamic framework. In terms of usability, we verified that the models had an adequate annotation, *i.e.* balanced equations, intuitive metabolite and reaction names, compartmentalization, functional gene-reaction associations and adequate representation of the central metabolism, among others. We then evaluated model similarity to experimental data from two chemostats (Table 1) using the normalized square differences between experimental and simulated rates (Equation 1):

$$F_i = \sum_{j=1}^{2} \frac{1}{n_j} \cdot \sum_{k=1}^{n_j} \frac{\sqrt{\left(v_{exp_{k,j}} - v_{mod_{k,j}}\right)^2}}{v_{exp_{k,j}}} \quad (1)$$

Here, F is the overall fitting relative error of model i, n_J corresponds to the number of predicted rates determined in each dataset (12 in dataset 1 and 30 in dataset 2). Also, v_{exp_k} corresponds to the vector of experimental rates of condition k in dataset j and $v_{mod_{k,j}}$ is the model's estimation of the experimental rates of condition k in dataset j.

For each prediction, we first constrained each model with n_J-1 experimental rates. Then, Flux Balance Analysis (FBA) (Orth, Thiele, & Palsson, 2010) was performed using biomass maximization as objective function to predict the remaining one.

It is worthy to note that whenever a model yielded an infeasible solution (due to carbon imbalance) or erroneously predicted the production of a compound under certain experimental condition, an error of 100% was assumed for that particular rate.

The model that gave best predictions compared to experimental data was chosen as the basis for the dynamic model. We tested four genome-scale metabolic models of *Pichia pastoris* that were available at the beginning of this study: the iPP669(B. K. S. Chung et al., 2010), the iLC915(Caspeta et al., 2012), the PpaMBEL1254(Sohn et al., 2010) and an updated version of the iPP669 model called iFS670. This model included the D-Arabitol (Cheng et al., 2014) synthesis pathway and equations for the heterologous synthesis of a FAB fragment, Human Serum Albumin (HSA) and thaumatin, according to *P. pastoris* codon usage (De Schutter et al., 2009) (Supplementary material 1). We also performed this analysis with the latest *Pichia pastoris* GSM published in January 2016 (Tomàs-Gamisans et al., 2016), to determine if it is worth including it in future versions of the dynamic framework (Supplementary material 11).

Table 1 - Chemostat data used for model selection

Set	Type of data	Rates	Conditions	Reference
1	Glycerol- and/or methanol-limited chemostats	5	4	(Solà et al., 2007)
2	Glucose-limited chemostats at different oxygen levels	7	6	(Carnicer et al., 2009)

4.2. Model construction

The structure of the model was based on an existing dFBA framework developed by Sanchez *et al* for *Saccharomyces cerevisiae* (Sánchez, Pérez-Correa, & Agosin, 2014). The model operates under a pseudo-steady state assumption (Mahadevan et al., 2002; Stephanopoulos, Aristidou, & Nielsen, 1998), i.e. considering that intracellular fluxes are several orders of magnitude faster than extracellular rates and, therefore, the former can be disregarded if the FBA model is resolved iteratively in short integration periods.

The model is composed by three linked blocks that are solved iteratively; (i) the kinetic block, (ii) the metabolic block and (iii) the dynamic block ().

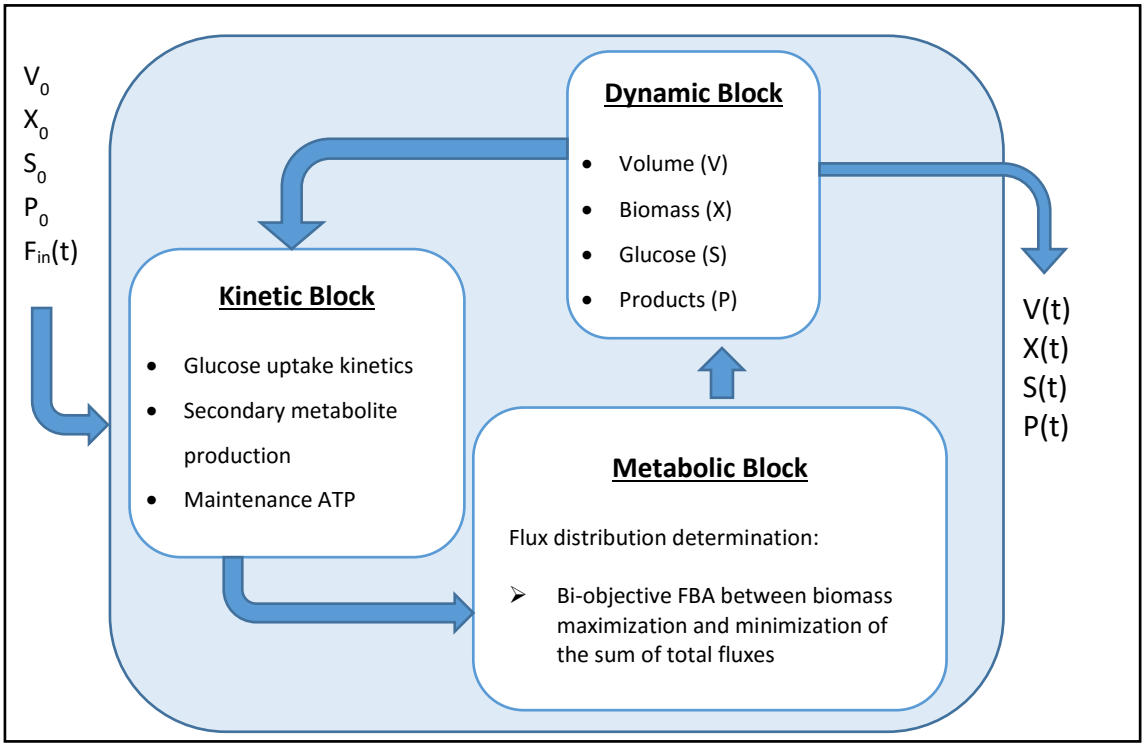


Figure 2 - Iterative structure of the model. V refers to culture volume [L], F_{IN} is the feeding policy used in fed-batch cultures, X, S and P are biomass, limiting substrate and Product concentration in [g/L] respectively.

Kinetic block

The kinetic block sets the uptake and production rates for all the compounds in the model. First, glucose uptake rate (v_{Gluc}) is determined using Michaelis-Menten kinetics (Postma, Verduyn, Scheffers, & Van Dijken, 1989).

$$v_{Gluc} = \frac{v_{S\,Max} \cdot S}{K_S + S} \qquad (2)$$

Here, S is the glucose concentration in the medium [g/L], $v_{S\,Max}$ is the maximum glucose uptake rate and K_S is the uptake half activity constant of this substrate. Once determined, the glucose uptake rate [mmol/g_{DCW}·h^{-1}] is included as lower bound of the corresponding exchange reaction in the model, which carries a negative flux if glucose is consumed.

Then, the lower bounds of the exchange reactions of the remaining compounds are fixed.

$$lb_k = v_{P_k} \qquad k = 1 \ldots 4 \qquad (3)$$

Where lb_K is the lower bound of the exchange reaction of compound k, which refers to the four products included: ethanol, pyruvate, arabitol and citrate. These parameters are redefined during the fed-batch phase; therefore, they adopt two values during this type of cultivation.

Finally, this block fixes the non-growth associated maintenance ATP (m_{ATP}, flux through a cytosolic ATP hydrolysis reaction in the model), which accounts for the energy drain caused by cellular processes not related to generating new cell material such as osmoregulation, shifts in metabolic pathways, cell motility etc. (Van Bodegom, 2007; Varela, Baez, & Agosin, 2004)

Metabolic block

The metabolic block receives a constrained metabolic model from the kinetic block and solves an optimization problem to determine growth rate and the flux distribution in the cell. The Genome Scale Model (GSM) consists on a set of *m* metabolites and n reactions grouped in a Stoichiometric Matrix S (m x n) that represents the cell's entire metabolism.

If accumulation of metabolites is neglected, a mass balance can be stated according to the equation (4):

$$S \cdot v = 0 \qquad (4)$$
$$s.t.$$
$$lb < v < ub$$

Where v is a vector of metabolic fluxes in [mmol/g_{DCW}·h]. Additionally, lower and upper bounds ($lb\ and\ ub$) for each component of the flux vector can be stated according to reaction reversibility, along with an objective function to solve the underdetermined system.

The metabolic block solves a bi-objective Quadratic Programming (QP) problem between maximization of growth rate and minimization of the total absolute sum of fluxes (Holzhütter, 2004), subjected to the constraints imposed by the stoichiometric matrix (Feng, Xu, Chen, & Tang, 2012):

$$Min\ \alpha \cdot \sum_{i=1}^{n} v_i^2 - (1-\alpha) \cdot \mu$$
$$s.t. \qquad (5)$$
$$S \cdot v = 0$$
$$lb_i \leq v_i \leq ub_i \qquad i = 1 \ldots n$$

In this formulation, the sub-optimal growth coefficient α is an adjustable parameter from the model and is used to modulate the importance of the two, biologically relevant, competing objectives (Sánchez, Pérez-Correa, et al., 2014; Schuetz, Kuepfer, & Sauer, 2007; Schuetz, Zamboni, Zampieri, Heinemann, & Sauer, 2012).

All Flux Balance Analyses (FBA) were solved using the Constraint-Based Reconstruction and Analysis (COBRA) Toolbox (Becker et al., 2007; Hyduke et al., 2011), which employs the programming library libSBML (Bornstein, Keating, Jouraku, & Hucka,

2008) and the SBML Toolbox(Keating, Bornstein, Finney, & Hucka, 2006). Finally, Gurobi 6.0.2 was used as optimization software.

Dynamic block

The dynamic block consists of a set of ordinary differential equations (ODEs) that account for the volume change of the culture and the mass balances of biomass and the dissolved compounds considered in the model in either batch or fed-batch configuration.

$$\frac{dV}{dt} = F(t) - SR \quad (6)$$

$$\frac{d(VX)}{dt} = \mu \cdot (VX) \quad (7)$$

$$\frac{d(VS)}{dt} = F(t) \cdot S_F - v_S \cdot MW_G \cdot (VX) \quad (8)$$

$$\frac{d(VP_k)}{dt} = v_{P_k} \cdot MW_{P_k} \cdot (VX) \quad (9)$$

Where V is volume [L], t is time [h], F(t) is the feed function for the fed-batch phase in [L/h] and SR is a constant sampling rate [L/h] determined for every fed-batch cultivation included. This was considered to simulate accurately the quantity of substrate added in the feed phase: during the batch phase, we took between 15 to 20% of the reactor volume in samples and the remaining volume was considered in the determination of the feeding rate. X is the biomass concentration [g/L], μ is the specific growth rate [h^{-1}] (obtained from equation 5), S is the extracellular concentration glucose [g/L], S_F is the feed's glucose concentration [g/L], P_K is the k-th extracellular product concentration in [g/L], v_{P_k} is the corresponding production rate [mmol/g_{DCW}·h] and MW accounts for the corresponding molecular weight [g/mmol]. All cultivations had glucose as the limiting substrate, and ethanol, pyruvate, arabitol and citrate as the main metabolic products. Therefore, Equation 9 comprises four differential equations.

The set of equations was solved in Matlab and solver options were the same as the ones used by Sanchez *et al* (Sánchez, Pérez-Correa, et al., 2014).

Model parameters

All the parameters studied, along with their units, lower (LBs) and upper (UBs) bounds and initial values for all optimizations are summarized in (Table 2). The LBs and UBs of $v_{S\,Max}$, K_S, and m_{ATP} were chosen according to the literature (B. K. S. Chung et al., 2010; van Urk, Postma, Scheffers, & van Dijken, 1989; Villadsen et al., 2011) while the rest of the bounds were selected to ensure that the algorithm had enough search space (upper bounds exceeded previously reported values of production and consumption rates (Carnicer et al., 2009)). Finally, initial values for parameter estimation were chosen to attain an initial feasible simulation.

Table 2 - Parameters of the model. In this table we present the symbol, name, units, lower bounds, initial value and upper bounds of the parameters.

Symbol	Name	Units	LB	Initial value	UB
$v_{S,max}$	Maximum glucose uptake rate	$mmol/g_{DCW}h$	0	2.5	10
K_S	Half saturation constant for glucose uptake	g/L	0	10^{-4}	10^{-3}
$v_{EtOH,B}$	Ethanol minimum secretion rate (batch)	$mmol/g_{DCW}h$	0	0.5	3
$v_{Pyr,B}$	Pyruvate minimum secretion rate (batch)	$mmol/g_{DCW}h$	0	0.1	2
$v_{Arab,B}$	Arabitol minimum secretion rate (batch)	$mmol/g_{DCW}h$	0	0.2	2
$v_{Cit,B}$	Citrate minimum consumption rate (batch)	$mmol/g_{DCW}h$	0	0	2
$v_{EtOH,FB}$	Ethanol minimum consumption rate (fed-batch)	$mmol/g_{DCW}h$	0	0	2
$v_{Pyr,FB}$	Pyruvate minimum consumption rate (fed-batch)	$mmol/g_{DCW}h$	0	0	2
$v_{Arab,FB}$	Arabitol minimum consumption rate (fed-batch)	$mmol/g_{DCW}h$	0	0	2
$v_{Cit,FB}$	Citrate minimum consumption rate (fed-batch)	$mmol/g_{DCW}h$	0	0	2
α_B	Sub-optimal growth coefficient (batch)	$[-]$	0	0	10^{-3}
α_{FB}	Sub-optimal growth coefficient (fed-batch)	$[-]$	0	0	10^{-3}
m_{ATP}	Non-growth associated ATP	$mmol/g_{DCW}h$	0	2	10
T_{Fed}	Time when secondary metabolite consumption starts in fed-batch cultures	h	20	25	32

4.3. Calibration with experimental data

4.3.1. Strains

Four *Pichia pastoris* strains were employed in this study: a parental GS115 strain (Invitrogen) and three recombinant strains harboring one, five and eight copies of the Thaumatin gene respectively. The strains were constructed and gently facilitated by the PhD candidate Alexandra Lobos. Despite the strains were transformed, thaumatin was not detected at concentrations higher than 100 ug/L in the cultivations. Therefore, its production was left out of the analysis and no mass balance was established to it.

4.3.2. Experiments

The batch model was calibrated using eight aerobic glucose limited cultivations corresponding to duplicates of the four strains available. On the other hand, the fed-batch configuration of the model was calibrated with three cultures of the strain harboring one copy of the gene under the same environmental conditions of the batch cultivations.

4.3.3. Cultivation Conditions

Each culture started from a 2 [mL] cryotube of the corresponding strain kept at -80 °C. A pre-culture was grown overnight at 30 °C in shake flasks with 50 [mL] of the inoculum medium until it reached 1 OD_{600}, which were then added to 450 [mL] of the batch medium to reach an initial volume of 500 [mL] and 0.1 initial OD_{600} (in both batch and fed-batch experiments). Culture conditions were kept at 30 °C, pH = 6.0 and DO 2.8 [mg/L]. Aerobiosis was achieved by a triple split-range control action including agitation (200–800 [RPM]), air flow (0.3–1.2 [L/min]) and pure oxygen flow (0–1.2 [L/min]) (M. Cárcamo et al., 2014). pH was controlled using phosphoric acid 20% [v/v] and ammonium hydroxide 20% [v/v]. Temperature was controlled with a mixture of hot and cold water, using the glass jacket of the reactors. Lastly, foam was controlled manually using silicone antifoam 10% [v/v]. Glucose starvation was detected when a sudden decrease of the CO_2

composition in the off-gas occurred, and confirmed each time using Benedict's reagent. For fed-batch experiments, the feed F(t) was designed to track a predefined time variable growth rate and, therefore, can be calculated from the reactor's glucose and biomass mass balances, as detailed elsewhere (Villadsen & Patil, 2007):

$$F(t) = \frac{\mu_{set(t)}}{S^F \cdot Y_{SX}} \cdot V_i X_i \cdot \exp\left(\int_{t_i}^{t} \mu_{set}(t)dt\right) \quad (10)$$

with S^F the glucose feed concentration [g/L], Y_{SX} the experimental glucose-biomass yield [g_{DCW}/g], t_i the time at which the feed started for a given cultivation [h], V_i and X_i the volume [L] and biomass [g/L] values at t_i, respectively, and $\mu_{SET}(t)$ the variable growth rate. The latter was defined as follows:

$$\mu_{set(t)} = (\mu_{max} - \mu_{min}) \cdot e^{-Ct} + \mu_{min} \quad (11)$$

Where μ_{MAX} = 0.1 [1/h], μ_{MIN} = 0.07 [1/h] and C = 0.07 [1/h]. Therefore, $\mu_{SET}(t)$ decays exponentially from 0.1 to 0.07 [1/h], which has been found to increase (in contrast to constant growth rates in the feed phase) the final biomass concentration in fed-batch cultivations of *E. coli* and *S. cerevisiae* performed in our laboratory (Martín Cárcamo, 2013).

4.3.4. Culture media

The components of the different media used in this study are detailed in Table 3, they were based on the recipe from (Tolner, Smith, Begent, & Chester, 2006). Sodium hydroxide was added until reaching a pH of 6 and the trace element solution was also taken from (Tolner et al., 2006).

Table 3 - Composition of 1L of the different define media used in this study

			Batch model development	**Fed-batch model development**	
Element	**Unit**	**Inoculum**	**Batch**	**Batch**	**Feed**
Glucose	[g]	10	50	40	500
$(NH_4)_2SO_4$	[g]	1.8	9	7.2	-
$MgSO_4 \cdot 7H_2O$	[g]	2.3	11.7	9.3	9
K_2SO_4	[g]	2.9	14.7	11.7	-
Histidine	[mg]	80	400	320	4
Sodium hexametaphosphate	[g]	5	25.1	20	-
Biotin	[mg]	0.32	1.6	1.3	100
Trace elements	[ml]	0.8	4	3.2	12.5

4.3.5. Bioreactor setup

The fermenters employed for all cultivations consisted of in-house built 1 L bioreactors (glass purchased from Garg Scientific, India). The main features of the setup, probes, gas detectors and peristaltic pumps are the same as the ones used in Sánchez *et al* (Sánchez, Pérez-Correa, et al., 2014).

4.3.6. Analytical procedures

Sampling and biomass determination

Samples of ~6 mL were periodically collected (every 2-3 hours) from all fermentations. Biomass was measured in OD at 600 nm using an UV-160 UV-visible recording spectrophotometer (Shimadzu, Japan). Biomass concentration was determined using the linear relationship: 1 OD_{600} = 0.72 [g/L], obtained using the method from (Marx,

Mecklenbräuker, Gasser, Sauer, & Mattanovich, 2009). Then, samples were centrifuged at 10000 rpm for 3 minutes and the supernatant stored at -80°C for further analysis.

Extracellular metabolite concentration determination

Glucose, ethanol, arabitol, citrate and pyruvate extracellular concentrations were quantified in duplicate by High Performance Liquid Chromatography (HPLC). Samples were prepared by adding 360 µL of culture supernatant (or a dilution of it), 40 µL of a 50 g/L pivalic acid solution (used as internal standard) and 0,1 µL of H_2SO_4 98% [v/v]. Then, 30 µL of the resulting solution were injected to a LaChrom L-7000 HPLC (Hitachi, Japan) equipped with an Aminex HPX-87H anion-exchange column (Bio-Rad, USA) for organic acids, alcohols and sugars separation, working at 35 °C with a 0.45 [mL/min] flow of the 5 [mM] H_2SO_4 mobile phase. A LaChrom L-7450A diode array detector (Hitachi, Japan) was set at 210 [nm] for detecting organic acids, and a LaChrom L-7490 refraction index detector (Hitachi, Japan) for sugars and alcohols. Finally, each metabolite was quantified normalizing the corresponding area in the chromatogram with the internal standard area and employing an external standard curve.

4.3.7. Objective Function of Model calibration

For model calibration, we formulated a nonlinear programming problem (NLP) with the dFBA embedded as a constraint. This optimization problem minimized the sum of square errors between the experimental data and the simulation output by searching the parameter space using the enhanced scatter search algorithm (eSS) (Egea & Balsa-Canto, 2009), which has been successfully used to solve bioprocessing problems (Balsa-Canto, Rodriguez-Fernandez, & Banga, 2007; Sacher et al., 2011; Sriram, Rodriguez-Fernandez, & Doyle, 2012). The objective function F used in the minimization was also normalized by the maximum corresponding measured variable in order to give all data a similar weight, regardless of their order of magnitude.

$$F = \min_{\theta} \sum_{i=1}^{m} \sum_{j=1}^{n} \left(\frac{X_{ij}^{mod} - X_{ij}^{exp}}{\max_{j} \left(X_{ij}^{exp} \right)} \right)^2 \quad (12)$$

With θ representing the parameter space, m the number of measured variables, n the number of measurements per variable, X_{ij}^{mod} the dFBA output of variable i and measurement j, X_{ij}^{exp} the corresponding experimental value and $\max_{j}(X_{ij}^{exp})$ the maximum value measured for variable i.

4.4. Reparametrization

4.4.1. Pre/post regression diagnostics

First, we briefly explain the regression diagnostics performed in this study, which are the basis of the reparametrization algorithm (Sacher et al., 2011). This section was taken from Sánchez *et al* (2014).

Sensitivity corresponds to the impact of the parameters on the state variables of the model. Here, we calculated the relative sensitivity of parameter k on state variable i (G_{ik}) as:

$$G_{ik}(t, \theta_k) = \frac{\theta_k}{X_i(t)} \cdot \frac{dX_i(t)}{d\theta_k} \quad (13)$$

Where t is time, $X_i(t)$ is the *i*th state variable in time t and θ_k is the *k*th parameter. With all G_{ik} values, we formed a sensitivity matrix G(t) for each experimental time, in which the *k*th column denotes the sensitivity of the kth parameter on the state variables. In order to obtain a single normalized score (spanning all experimental times) of each parameter over each variable, we calculated average sensitivities as detailed in Hao *et al* (Hao, Zak, Sauter, Schwaber, & Ogunnaike, 2006). Therefore, if this score is under 0.01 in each variable for a given parameter, we chose to fix the corresponding parameter.

To calculate identifiability, the MATLAB function *corrcoef* was used to determine the correlation coefficients between each column of the sensitivity matrices, and stored the information in a correlation coefficients matrix (C). If any of the off-diagonal elements of the matrix take on values over a certain threshold (in our case $|Cij| \geq 0.95$), it is considered that the associated parameters are strongly correlated, and therefore one of them should be fixed.

To determine parameter significance, and also using the sensitivity matrices, we first calculated the Fisher Information Matrix (FIM) (Petersen, Gernaey, & Vanrolleghem, 2001):

$$FIM = \sum_{j=1}^{n} G_j^T Q_j G_j \qquad (14)$$

Here, Gj is the sensitivity matrix for measurement j, n is the number of measurements, and Qj is the inverse of the measurement error covariance matrix assuming white and uncorrelated noise, which is used as a weighting matrix. Using this matrix, the variances for each estimated parameter (σ_k^2) were calculated as (Landaw & DiStefano 3rd, 1984; Petersen et al., 2001)

$$\sigma_k^2 = FIM_{kk}^{-1} \qquad (15)$$

With the variances we computed the confidence interval (CI) with 5% significance for the *k*th parameter as follows:

$$CI_k = \left[\hat{\theta}_k \pm 1.96\sigma_k\right] \qquad (16)$$

Here, $\hat{\theta}_k$ is the estimated value of the corresponding parameter. Finally, coefficients of confidence (CC) were calculated as follows:

$$CC_k = \frac{\Delta(CI_k)}{\hat{\theta}_k} = \frac{3.92\sigma}{\hat{\theta}_k} \qquad (17)$$

With $\Delta(CI_k)$, the CI's length. With this metric, we determined that a parameter was not significantly different from zero if the CI contained the zero, therefore if the absolute value of the CC was equal or larger than 2.

4.4.2. General approach

A reparametrization procedure called HIPPO (Sánchez, Soto, Jorquera, Gelmi, & Pérez-Correa, 2014) (Heuristic Iterative Procedure for Parameter Optimization, http://www.systemsbiology.cl/tools/) was applied to address the parametric problems in the model.

HIPPO started by performing sensitivity and identifiability tests on the initial calibration results for each dataset. Then, model parameters were iteratively fixed depending if they presented sensitivity and/or identifiability problems in the initial regression. Here, parameters were fixed one by one until the non-fixed subset presented none of the aforementioned problems. Finally, significance was determined for the remaining parameter set, also called reduced model structure, and if all parameters were significantly different from zero it was considered to be an *a priori* robust candidate for Cross calibration.

Due to the heuristic nature of HIPPO, several robust modeling structures were achieved for each experimental dataset (check (Sánchez, Pérez-Correa, et al., 2014) for details on the heuristic employed by the algorithm). These were then ranked using a score called Mean Coefficient of Confidence (MCC), which is the average of the Coefficients of Confidence of each of the parameters of the model:

$$MCC_i = \frac{\sum_{k=1}^{n_i} CC_k}{n_i} = \frac{\sum_{k=1}^{n_i} \left(\frac{\Delta CI_{k,5\%}}{\hat{\theta}_k} \right)}{n_i} = \frac{1}{n_i} \cdot \sum_{k=1}^{n_i} \frac{3.92 \cdot \sigma_k}{\hat{\theta}_k} \quad (18)$$

Where MCC_i is the mean coefficient of confidence for dataset i, CC_K the coefficient of confidence of parameter k, n_i the number of parameters of the model used to fit dataset i and $\Delta CI_{k,5\%}$ the length of the 95% Confidence Interval for the parameter k.

The modeling structure of each dataset with the lowest MCC was utilized as candidate for the cross-calibration stage.

4.5. Cross Calibration of available datasets using candidate robust modeling structures derived from the reparametrization stage.

After the reparametrization of each dataset, a candidate robust modeling structure was achieved. The latter was employed to recalibrate the rest of the datasets in order to evaluate its robustness. It is worthy to note that the parameters left out of the calibration were either fixed according to values reported in literature, assumed to be zero or fixed at the mean value achieved in the calibrations. This was done in order to avoid assuming a minimum production of compounds in batch cultivations and to ensure model convergence for parameters that had no reported values in literature (feed phase consumption rates) (Table 4). We applied a maximum of 2500 iterations in the scatter search algorithm.

The reduced modeling structures were evaluated according to four parameters:

I. **Relative difference between objective functions (F_{DIFF}),** corresponds to the average relative difference between the objective function achieved with a reduced modeling structure in contrast to the value achieved with the original model structure. This score was determined according to the following expression:

$$F_{DIFF} = \frac{1}{n} \cdot \sum_{i=1}^{n} \frac{F_{i,Reduced} - F_{i,Original}}{F_{i,Original}} \quad (19)$$

Where n corresponds to the number of cultures of each type, $F_{i,Original}$ is the fit objective function for dataset i using the original model structure and $F_{i,Reduced}$ is

the fit objective function of dataset i using a reduced, *a priori* robust, modeling structure.

Table 4 - Values at which problematic parameters were fixed in the cross calibration stage. Parameters marked with '-' in the reference column indicates that no *a priori* value was assumed for that particular parameter, which is the case for the batch minimum secretion rates. '*' meant that the value of a particular parameter was fixed at the mean value achieved in the calibrations, because no information about them could be found in literature. It is worthy to mention that fixing these parameters at zero, allowed no consumption of batch by-products and yielded poor fed-batch fittings (data not shown).

Parameter	Fixation Value	Units	Reference
$v_{S,max}$	6	$mmol/g_{DCW}h$	(van Urk et al., 1989)
K_S	0.0027	g/L	(van Urk et al., 1989)
$v_{EtOH,B}$	0	$mmol/g_{DCW}h$	-
$v_{Pyr,B}$	0	$mmol/g_{DCW}h$	-
$v_{Arab,B}$	0	$mmol/g_{DCW}h$	-
$v_{Cit,B}$	0	$mmol/g_{DCW}h$	-
$v_{EtOH,FB}$	1.21	$mmol/g_{DCW}h$	*
$v_{Pyr,FB}$	0.14	$mmol/g_{DCW}h$	*
$v_{Arab,FB}$	0.15	$mmol/g_{DCW}h$	*
$v_{Cit,FB}$	0.008	$mmol/g_{DCW}h$	*
α_B	0	$[-]$	(Morales et al , 2014)
α_{FB}	0	$[-]$	(Morales et al., 2014)
m_{ATP}	2.18	$mmol/g_{DCW}h$	(B. K. S. Chung et al., 2010)
T_{Fed}	22	h	*

II. **Percentage of Significance issues** refers to the number of times a parameter was found to be not significantly different from zero out of the total of significance determinations performed for a particular structure. If a model structure had 6 parameters and was used to calibrate 8 datasets, 48 significance determinations were performed for that particular model.

III. **Percentage of Sensitivity issues** refers to the number of times one of the estimated parameters showed no impact over state variables (average sensitivity < 0.01) for

a particular modeling structure out of the total sensitivity determinations performed.

IV. **Percentage of Identifiability issues**, corresponds to the number of times a pair of parameters presented a strong correlation (≥ 0.95), out of the total parameter pairs of a particular modeling structure. If p is the number of parameters of the model and n is the number of datasets used for calibration, the total of parameter pairs for which identifiability was determined is:

$$Total\ pairs = \frac{p \cdot (p-1)}{2} \cdot n \qquad (20)$$

Finally, the modeling structure that presented the lowest F_{DIFF} and fewest parametric problems was used as a candidate robust modeling structure for the corresponding type of culture.

4.6. Robustness check of the chosen modeling structure

Once a candidate robust modeling structure was determined for the batch and fed-batch cases, we tested its robustness (absence of parametric problems) by calibrating new experimental data. If the calibration resulted in a close fit to the data and presented no identifiability, sensitivity or significance problems, the model structure was considered like robust. For the batch model, we employed fermentation data from *P. pastoris* GS115 strain grown with 40 [g/L] of glucose as carbon source at T° = 25°C and pH = 6. The robustness of the fed-batch model was evaluated with a glucose-limited cultivation consisting of a 60 [g/L] of glucose batch phase and an exponential feed using 500 [g/L] of glucose. The medium was added in the feeding phase in order to achieve an exponentially decreasing growth rate from 0.1 to 0.07 [1/h].

4.7. Model validation

Finally, the predicting capability of the robust batch and fed-batch models was evaluated for conditions similar to the ones used in the initial calibrations (training set).

In the case of the robust batch model, we first calibrated the duplicates of the strain harboring one copy of the thaumatin gene together, obtaining a characteristic parameter set for the strain. Then, we used these parameters to predict the course of a different batch cultivation performed in the same conditions (30°C and pH 6).

This procedure was also applied for the fed-batch configuration. Here, we simulated a bioreactor dynamics using the parameters obtained in the best calibration of the training dataset (the one in which the calibration objective function was minimal compared to the rest of the calibrations) using the robust modeling structure chosen previously. This prediction was compared with experimental data of a different fed-batch cultivation.

4.8. Simulation

4.8.1. Analysis of the metabolic flux distribution during different stages of a dynamic cultivation

After the calibration of the dataset used to check the fed-batch model robustness, we extracted the flux distribution from three metabolically different stages of the cultivation in order to analyze *P. pastoris'* central metabolism:

i. Exponential growth during the batch phase
ii. Ethanol and arabitol consumption during glucose starvation
iii. Controlled growth during the feeding phase.

In addition, we checked flux directionality against reported values and modified the genome-scale metabolic model until we observed an agreement between experimental and

predicted fluxes. Finally, we recalibrated the data with the curated model and analyzed the flux distribution of *P. pastoris* central metabolism in the aforementioned stages.

4.8.2. Discovery of beneficial knock-out targets for the overproduction of recombinant Human Serum Albumin

To demonstrate potential applications of the model, gene targets for the overproduction of the recombinant protein Human Serum Albumin (HSA) were searched by simulating the growth and protein secretion of single knock-out strains of *P. pastoris* in batch cultivations. To do this, we included in the Metabolic Block a second quadratic programing problem to simulate the behavior of a knock-out strain. The second problem consisted in the Minimization of Metabolic Adjustment (MOMA) algorithm (Segrè, Vitkup, & Church, 2002), which states that, after a genetic perturbation, the cell will attempt to redistribute its metabolic fluxes as similar as possible to the parental strain (equation 22).

The hypothetical parental strain was characterized using the parameters obtained in the calibration of the dataset used for the batch model validation plus a reported specific HSA productivity (q_P) by *P. pastoris* growing in glucose (Rebnegger et al., 2014), which depended on the specific growth rate μ (). In every iteration of the model, the minimum HSA production was fixed according to this relation, which was assumed to be a third degree polynomial just of modeling purposes. Other kinetics might be used to represent the q_P vs μ relation, but this depends on the strain and protein being produced (Maurer et al., 2006).

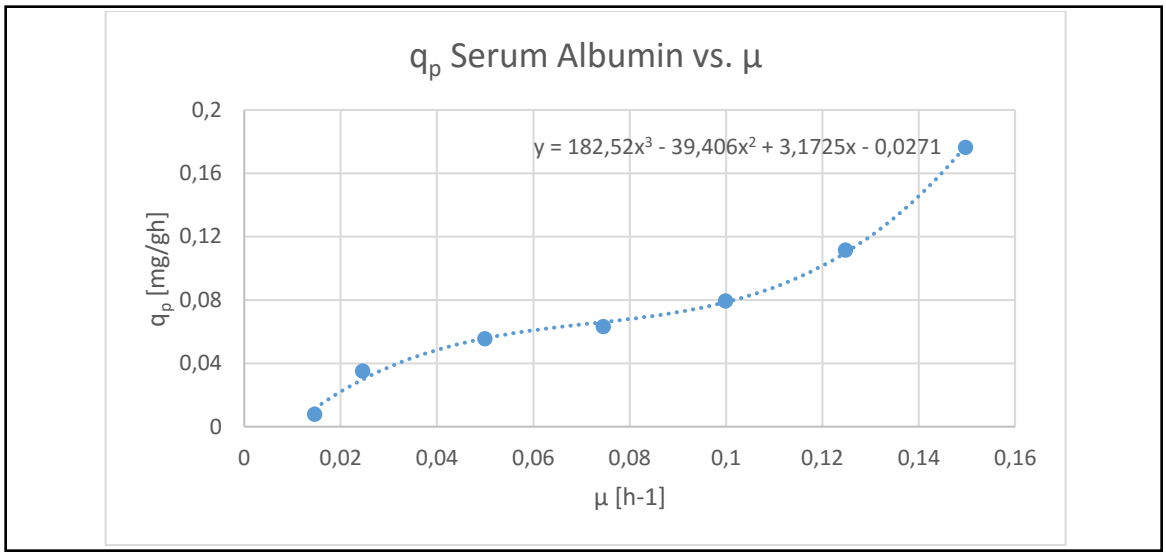


Figure 3 - Relation between Human Serum Albumin production and growth rate in a glucose limited chemostat taken from Rebnegger et al, 2014

Once the constrained model enters the metabolic block, it first solves equation 5, from which it obtains the parental flux distribution v_0. Then, the k reactions associated with gene j are blocked:

$$lb_{l,j} = ub_{l,j} = 0 \qquad l = 1 \dots k \tag{21}$$

Finally, the MOMA algorithm uses the flux distribution of the parental strain v_0 to calculate the knockout distribution v_{KO} as the Euclidean distance between them, considering that the actual model has the corresponding deletion.

MOMA:

$$Min \left(v_0 - v_{KO,j}\right)^2 \tag{22}$$

$$s.t.$$

$$lb_i \leq v_{i,KO,j} \leq ub_i \qquad i = 1 \dots n$$

We performed one batch simulation for every gene in the model, which were evaluated in terms of final protein and biomass concentrations and compared its performance against the parental strain. The candidates that reached a higher HSA concentration than the parental strain were manually analyzed and some of them were proposed as candidates to improve HSA production.

4.8.3. Evaluation of different feeding policies in silico to improve recombinant protein production considering specific information about the strain and process setup

We also evaluated feeding policies to improve the volumetric productivity of HSA production, considering information about a strain of interest and process limitations.

Simulations were performed using the parameters obtained in the calibration of the fed-batch validation dataset and including HSA biosynthesis in the mass balances. We used the same volumetric productivity (q_P) vs growth rate (μ) relation from (Rebnegger et al., 2014) to determine protein production as a function of the growth rate. The process limitations (based in our setup) were a maximum reactor volume of 1 L, and a maximum oxygen transfer rate of 10.9 [g/L·h], given by a k_La of 300 h^{-1} and a driving force of ($C_{O2,SATURATION} - C_{O2,SETPOINT}$) of 36.2 [mg/L], considering the incorporation of pure oxygen into the bioreactor. If any of these thresholds were violated by either the feeding rate of medium or the oxygen uptake rate (extracted from the model), the integration stopped.

We evaluated 13 exponential feeding policies. Five of them maintained a constant growth rate during the feeding phase and the rest considered a decreasing growth rate throughout the culture (Supplementary Material 3). After the simulation, we ranked the strategies according to the volumetric productivity of recombinant HSA and chose the best one as a culture alternative that could potentially improve bioreactor performance.

5. RESULTS AND DISCUSSION

5.1. Genome-Scale model selection from literature

The main components and the relevant usability features of published GSMs of *Pichia pastoris* are detailed in Table 5. The PpaMBEL1254 model was discarded due to the lack of intuitive reaction and metabolite names, as well as the absence of gene-protein relations (at least in the online version), hampering the analysis of knock-out strains. All the models share the same structure of the central metabolism, which carries most of the flux entering the cell.

Table 5 - Main components and usability features of available genome-scale metabolic models of *Pichia pastoris*

	iPP669	**iFS670**	**PpaMBEL1254**	**iLC915**
Number of genes	669	670	540	915
Reactions	1354	1383	1254	1426
Metabolites	1177	1195	1058	1302
Compartments	8	8	8	6
Platform used for analysis	Cobra	Cobra	Cobra	Raven
Intuitive nomenclature for reactions and metabolites	Yes	Yes	No	No
Capable of performing Single Gene deletions	Yes	Yes	No	Yes
Capable of automatically checking mass balance	No	No	No	Yes

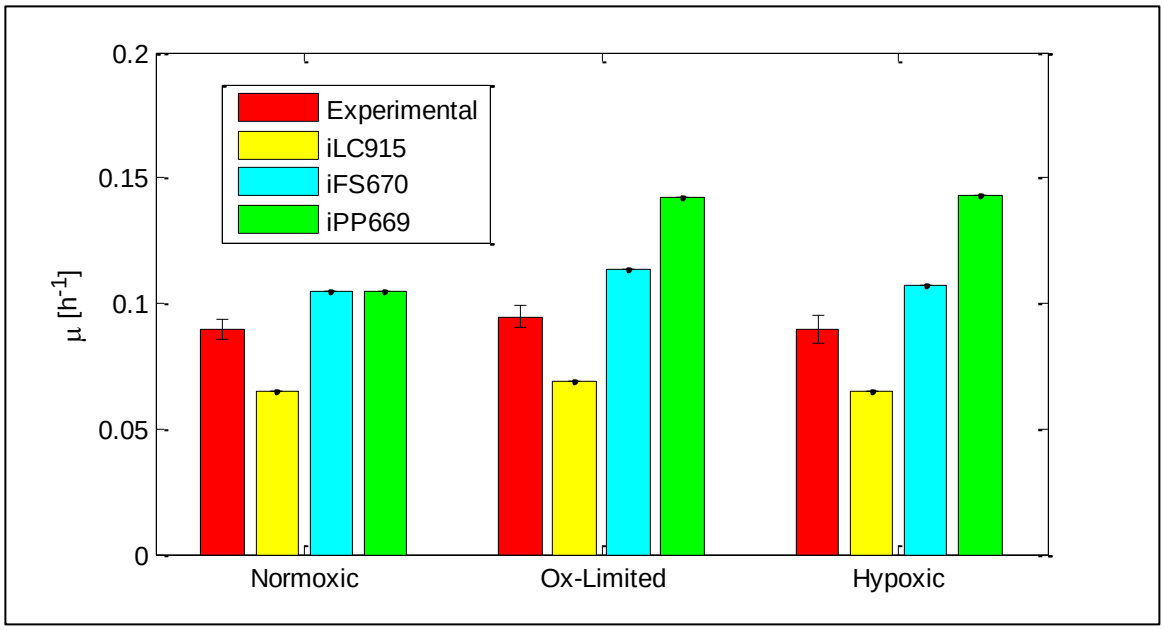
After the determination of the average relative error between model predictions and experimental data (Carnicer et al., 2009; Solà et al., 2007) (Table 6), we selected the iFS670 model since it has a desirable structure and better reproduces experimental data from *P. pastoris* chemostats. It is worth mentioning that the inclusion of the arabitol

biosynthesis pathway into Chung's (iPP669) model – resulting in the iFS670 model - greatly improved the predictions of specific growth rate, Oxygen Uptake Rate (OUR) and Carbon Dioxide Evolution Rate (CER) in hypoxic glucose-limited chemostats (Supplementary Material 1). Specifically, the deviation of carbon towards arabitol reduced the predicted growth rate in those conditions when compared to the iPP669 model (Figure 4), resulting in a reduction of the difference with the corresponding experimental value.

Table 6 - Average error of model predictions using two datasets from carbon-limited chemostats. In glycerol- and/or methanol (MetOH) – limited chemostats, the models were employed to predict specific growth rate μ, Oxygen Uptake Rate (OUR) and carbon dioxide evolution rate (CER) in four different conditions, which gives a total of 12 predictions. In the glucose limited chemostats, the models were used to estimate μ, OUR, CER, ethanol secretion rate and arabitol secretion rate in six conditions, which gives a total of 30 model predictions. Experimental data was taken from (Carnicer et al., 2009; Solà et al., 2007)

Carbon Source	**iLC915**	**iFS670**	**iPP669**	**Number of predictions**
Glycerol/MetOH	78%	37%	38%	12
Glucose	85%	36%	52%	30
Overall Error (F)	83%	36%	48%	42

Figure 4 - Experimental and model-predicted specific growth rates using glucose as the only carbon source at different oxygen levels for a *P. pastoris* wild type strain. Data taken from (Carnicer et al., 2009)



5.2. Batch model development

In this section, we present the steps followed to achieve a robust dFBA modeling structure of aerobic batch cultures of *Pichia pastoris*, capable of fitting new data with as few parametric problems as possible and predicting bioreactor dynamics.

5.2.1. Initial calibration

The initial structure of the batch model (eight parameters) was capable of fitting different cellular dynamics from eight aerobic batch cultivations (Figure 5 and Supplementary Material 5).

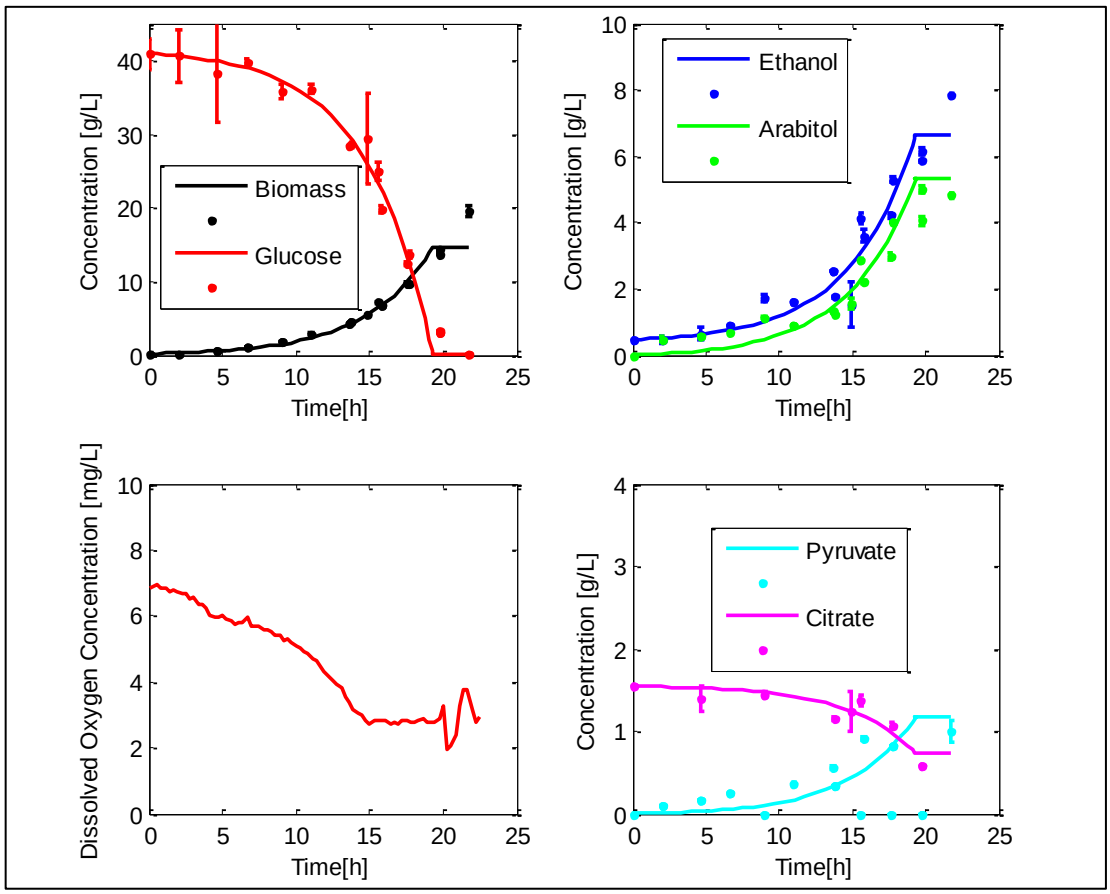


Figure 5 - Model calibration of a glucose-limited aerobic batch cultivation of *Pichia pastoris*. Experimental data from two replicates is shown with points and the model fit is presented in continuous lines. Also, the dissolved oxygen profile (in [mg/L]) is included in the bottom left graph.

Production of ethanol and arabitol was detected in the cultivations, which was probably caused by the high initial glucose concentration used in the experiments. Ethanol production during batch cultivations has already been reported (Heyland et al., 2010), however, the formation of these compounds has usually been associated to oxygen scarcity in glucose-limited conditions (Baumann et al., 2010; Carnicer et al., 2009).

The mean, minimum and maximum values of the model parameters, calculated for the calibrations, are presented in Table 7. Results show that the initial fittings covered a wide range of values. Here, the suboptimal growth parameter alpha (α_B) was always greater than zero, which indicates that the model "prefers" to include the minimization of total fluxes in the objective function, rather than only maximizing specific growth rate. This forces the metabolic block to solve a QP problem, which has the practical benefit of eliminating redundant cycles in the resulting flux distributions, also called Type III Pathways (Price, Famili, Beard, & Palsson, 2002), - which make metabolic pathway analyses cumbersome.

Table 7 - Minimum, Mean and Maximum parameter values achieved in batch model calibrations. More details on the calibrations can be found in Supplementary material 5.

Parameter	Minimum	Mean	Maximum	Units
V_{MAX}	1,27	4,286	7,948	[mmol/g_{DCW}·h]
K_S	1e-05	2,96E-04	9,80E-04	[g/L]
$v_{EtOH,B}$	0,024	1,363	2,968	[mmol/g_{DCW}·h]
$v_{Pyr,B}$	0,003	0,145	0,248	[mmol/g_{DCW}·h]
$v_{Arab,B}$	0,088	0,373	0,541	[mmol/g_{DCW}·h]
$v_{Cit,B}$	1e-05	0,177	1,104	[mmol/g_{DCW}·h]
α_B	1,454e-06	2,17E-04	4,15E-04	[-]
m_{ATP}	0,001	3,064	10,000	[mmol/g_{DCW}·h]

5.2.2. Parametric problems found in the initial batch model structure

Several parametric problems were found in the original modeling structure of batch cultivations. Table 8 shows how often a pair of parameters presented identifiability, sensitivity and significance issues in the eight datasets used in the initial calibrations. Non-growth associated maintenance ATP (m_{ATP}) was the parameter that presented the strongest correlations with other parameters in the model, such as maximum specific glucose uptake rate; ethanol and arabitol specific secretion rates; and suboptimal growth. This may result from the fact that a change in m_{ATP} affects directly the ATP-producing pathways in the metabolic model, resulting in changes in biomass and product yields, which are also defined by other parameters of the model.

Also, the glucose uptake saturation constant K_S was the only parameter with frequent sensitivity and significance problems, i.e. a potential candidate to be left out of the adjustable parameter set.

Table 8 - Parametric problems of the initial model structure for batch cultivation. Parameters were considered problematic if they presented a particular problem in more than 25% of the initial calibrations of the batch model (at least 3 out of 8). In the case of identifiability problems, we show the parameter pairs that recurrently presented a correlation higher than 0.95. The frequency (in percentage) of the different problems is shown in parenthesis.

Problem	Problematic Parameters
Identifiability	1. $v_{S\,Max}$ and m_{ATP} (63%) 2. $v_{S\,Max}$ and $v_{EtOH,B}$ (50%) 3. m_{ATP} and α_B (50%) 4. $v_{PYR,B}$ and $v_{EtOH,B}$ (38%) 5. m_{ATP} and $v_{EtOH,B}$ (38%) 6. m_{ATP} and $v_{Arab,B}$ (38%)
Sensitivity	1. K_S (38%)
Significance	1. K_S (50%)

5.2.3. Model reparametrization and cross calibration

After the calibration and determination of the parametric problems of each dataset, the non-relevant parameters were fixed (left out of the adjustable set) using the reparametrization procedure HIPPO. The remaining, *a priori* robust, reduced models (Table 9) were employed to recalibrate the available data (eight batch cultivations) in order to determine if they were enough to reproduce *P. pastoris* behavior appropriately. The persistence of parametric problems compared to the original model was also evaluated.

Table 9 - Potential Robust Structures Tested in the Cross-Calibration Stage. Each one of these structures was derived from the calibration of the corresponding dataset

Structure	Parameters included
Original	$V_{MAX}, K_S\ v_{EtOH,B}, v_{Pyr,B}, v_{Arab,B}, v_{Cit,B}, m_{ATP}\ and\ \alpha_B$
1	$V_{MAX}, v_{EtOH,B}, v_{Pyr,B}, v_{Arab,B}, v_{Cit,B}\ and\ \alpha_B$
2	$V_{MAX}, v_{Cit,B}\ and\ \alpha_B$
3	$K_S, v_{EtOH,B}, v_{Pyr,B}, v_{Arab,B}\ and\ v_{Cit,B}$
4	$v_{EtOH,B}\ and\ v_{Cit,B}$
5	$V_{MAX}, v_{Pyr,B}, v_{Arab,B}$
6	$V_{MAX}, v_{EtOH,B}, v_{Pyr,B}, v_{Arab,B}, v_{Cit,B}$
7	$V_{MAX}, K_S, v_{EtOH,B}, v_{Pyr,B}, v_{Cit,B}$
8	$K_S, v_{Pyr,B}, v_{Arab,B}, \alpha_B\ and\ m_{ATP}$

Structures 1 and 6 were the only parameter sets whose fitting capability was similar to the original model of eight parameters (Table 10), showing the importance of including the specific uptake and production rates of the compounds considered in the model. On the contrary, maintenance ATP and K_S were left out of these structures because of the frequent identifiability and sensitivity problems associated to them.

Structure 6 lacks the sub-optimal growth parameter α_B, which forces the solution of a linear programming (LP) problem of growth rate maximization in the metabolic block, since it was assumed to be zero if it was left out of the adjustable parameter set (See Table 4). This structure showed a significant improvement in significance and sensitivity issues compared to the original model; however, identifiability appeared to be a major problem (Table 10). Probably, the multiple solutions associated to an underdetermined LP problem may hamper the possibility of unambiguously infer the parameter values from the data. Identifiability is particularly important since the parameters are inputs (restrictions) for the calculation of flux distributions in genome-scale metabolic models.

Therefore, due to the recurrent identifiability issues found in structure 6, Structure 1 was preferred to calibrate a different dataset in order to validate it as the most robust modeling structure for aerobic, glucose-limited batch cultures of *Pichia pastoris*.

Table 10 – Batch Cross Calibration summary. Structures are organized according to F_{DIFF}, which corresponds to the average relative difference between the objective function achieved by the reduced model compared to the original. For more information on these scores refer to section 2.5. Structures that reduced the frequency of parametric problems are highlighted.

Structure	N° parameters	F_{DIFF}	Significance Issues (%)	Sensitivity Issues (%)	Identifiability Issues (%)
Original	**8**	**0**	**23.6**	**16.7**	**17.4**
1	6	-10%	22.9	18.8	15.0
6	5	+18%	0	2.5	61.3

5.2.4. Robustness check of the chosen reduced modeling structure for glucose-limited, aerobic batch cultures of *Pichia pastoris*

On this new dataset, Structure 1 showed a proper fit to the data and did not yield identifiability nor significance problems. However, the maximum glucose uptake rate ($v_{S\,Max}$) had no impact on the state variables. Therefore, after the initial calibration (not shown) we fixed it at 6 [mmol/g_{DCW}h]. illustrates the model fit and Table 11 presents the

parameter values and confidence intervals achieved in the second calibration, which had no identifiability, significance or sensitivity issues.

Despite the sensitivity issue associated with $v_{S\,Max}$ for this particular dataset, we included this parameter in the proposed robust modeling structure. This is because in the calibration of the cultures of the strain harboring eight copies of the thaumatin gene (cultures 7 and 8) the state variables were very sensitive to this parameter (average sensitivity > 0.7, recall that the sensitivity threshold is 0.01), and it was necessary to achieve a close fit to the data. Therefore, if this parameter is found insensitive in future calibrations, it could be easily fixed at reported values.

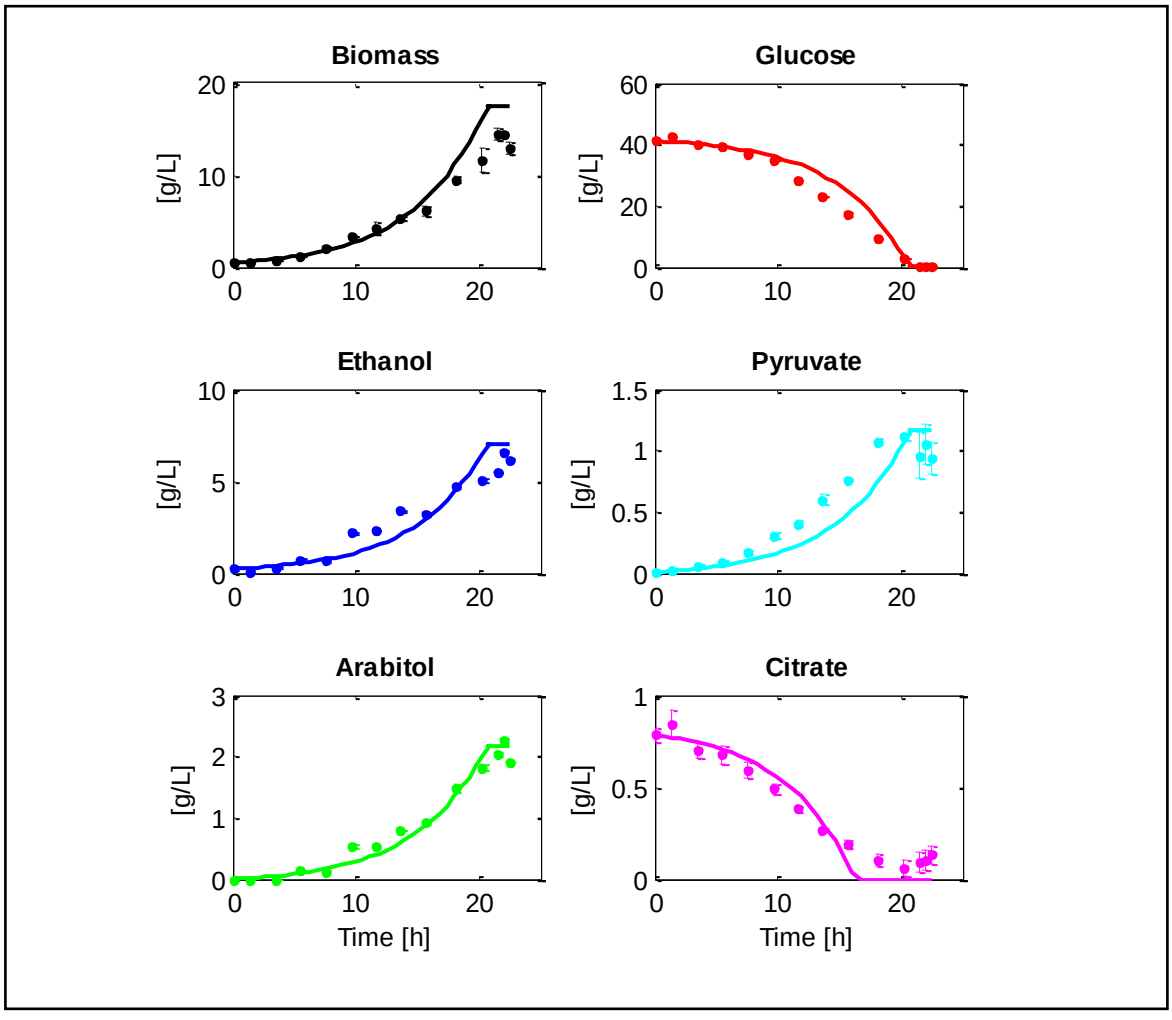


Figure 6 – Robustness check of Structure 1 as modeling framework for aerobic, glucose-limited batch cultures of *Pichia pastoris*. The figure shows the reduced model fit to the extracellular concentration of the species involved in the model during a batch cultivation.

Table 11 - Parameter values achieved in the validation of the batch model structure. Values of the parameters are presented together with their 95% confidence intervals. In this calibration, $v_{S\,Max}$ was fixed at a known value to avoid sensitivity issues. Finally, the calibration yielded no parametric problems.

Parameter	Value	Units
$v_{S\,Max}$	6	mmol/g_{DCW}·h
$v_{EtOH,B}$	1.47 ± 0.07	mmol/g_{DCW}·h
$v_{Pyr,B}$	0.13 ± 0.05	mmol/g_{DCW}·h
$v_{Arab,B}$	0.14 ± 0.06	mmol/g_{DCW}·h
$v_{Cit,B}$	0.09 ± 0.04	mmol/g_{DCW}·h
α_B	4.1 ± 0.9 ·10^{-4}	[-]

Summing up, we achieved a robust modeling structure for glucose-limited, aerobic batch cultivations of *Pichia pastoris*, composed of six parameters that estimate specific consumption and production rates of all the species involved in the mass balances. It also determines the specific growth rate by solving a bi-objective optimization problem, which avoids most of the identifiability issues arising between parameters.

5.2.5. Batch model validation

The parameters found for the strain harboring one copy of the thaumatin gene predicted correctly the dynamics of a different batch cultivation using the same strain (Figure 7). Biomass, glucose, ethanol, pyruvate and citrate dynamics were in good agreement with model simulations. On the other hand, arabitol formation was significantly overestimated, probably because in the training dataset the initial concentration of glucose was higher, which might have increased the formation of secondary products.

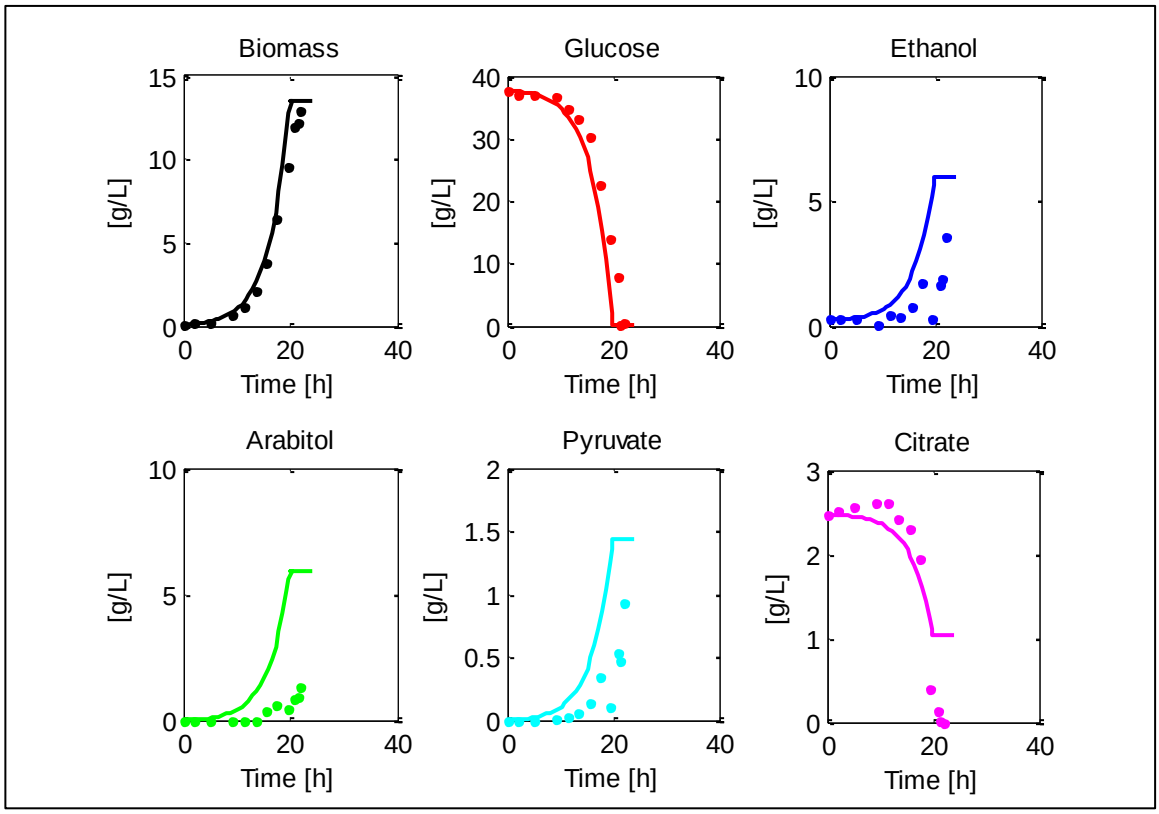


Figure 7 - Batch model preliminary validation. Data is shown in filled circles while model prediction is presented in continuous lines.

6.2. Fed-batch model development

6.2.1. Initial Model Calibration

Three aerobic, glucose-limited fed-batch cultivations where successfully calibrated with the initial model structure composed of fourteen parameters (Figure 8). The values of the parameters obtained in the three datasets can be found in Supplementary Material 7.

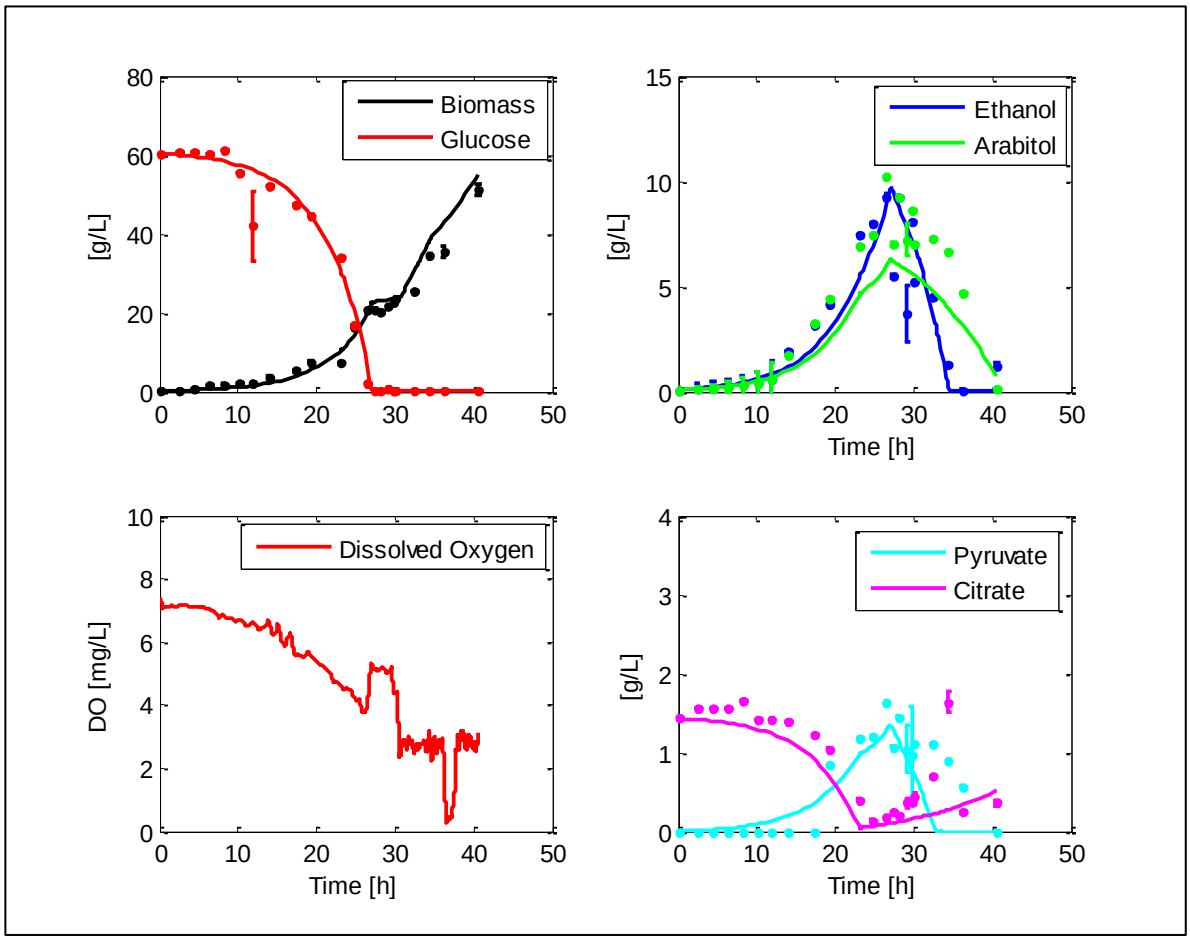


Figure 8 - Example of a model calibration of a fed-batch culture of *Pichia pastoris*. Experimental data with the corresponding standard deviation is shown in points while the model calibration is presented in continuous lines. The bottom left graph presents the dissolved oxygen concentration throughout the cultivation. The rest of the calibrations are included in Supplementary Material 7.

6.2.2. Parametric problems found in the initial fed-batch model structure

The initial fed-batch model also presented parametric problems. However, these were less recurrent in comparison with the batch model. The most frequent correlation (in two out of the three calibrations) was between the maximum glucose uptake rate and the ethanol formation rate during the batch phase (Table 12). Also, $v_{EtOH,B}$ and $v_{Arab,FB}$ parameters showed 5 and 6 strong correlations, respectively, with other parameters of the model.

Table 12 – Frequency (in %) with which a pair of parameters presented identifiability issues in the initial modeling structure of fed-batch cultures of *Pichia pastoris* (3 datasets). Parameters with recurrent identifiability issues are highlighted.

	$v_{S,max}$	K_S	$v_{EtOH,B}$	$v_{Pyr,B}$	$v_{Arab,B}$	$v_{Cit,B}$	$v_{EtOH,FB}$	$v_{Pyr,FB}$	$v_{Arab,FB}$	$v_{Cit,FB}$	α_B	α_{FB}	m_{ATP}	T_{Cons}
$v_{S,max}$														
K_S	0													
$v_{EtOH,B}$	67	0												
$v_{Pyr,B}$	0	0	0											
$v_{Arab,B}$	33	0	33	0										
$v_{Cit,B}$	0	0	33	0	0									
$v_{EtOH,Fb}$	0	0	33	0	0	0								
$v_{Pyr,FB}$	0	0	0	0	0	0	0							
$v_{Arab,FB}$	0	0	33	0	0	33	33	0						
$v_{Cit,FB}$	0	0	0	0	0	0	0	0	33					
α_B	0	0	0	0	0	0	0	0	33	0				
α_{FB}	0	0	0	0	0	0	0	0	0	0	0			
m_{ATP}	0	0	0	0	0	0	0	0	0	0	0	0		
T_{Cons}	0	0	0	0	0	0	0	0	33	0	33	0	0	

The citrate minimum secretion rate (both in batch and fed-batch phases) and the suboptimal growth coefficient during the feeding phase (α_{FB}) were the parameters that

presented more sensitivity and significance issues (Table 13). Also, α_B, K_S and the specific pyruvate consumption rate $v_{PYR,FB}$ were non-significant in at least one of the cultivations.

Table 13 - Percentage of times a parameter of the model presented sensitivity or significance problems out of a total of three model calibrations. Parameters with sensitivity or significance issues are highlighted.

	$v_{S,max}$	K_S	$v_{Et,B}$	$v_{Py,B}$	$v_{Ar,B}$	$v_{Ci,B}$	$v_{Et,FB}$	$v_{Py,FB}$	$v_{Ar,FB}$	$v_{Ci,FB}$	α_B	α_{FB}	m_{ATP}	T_{Cons}
Sensitivity	0	0	0	0	0	0	0	0	0	33	0	67	0	0
Significance	0	33	0	0	0	0	0	33	0	100	67	100	0	0

6.2.3. Model reparametrization and cross calibration

Once the parametric problems were determined after the initial calibration, we employed the reparametrization algorithm HIPPO to find subsets of parameters that could be sufficient to represent the dynamics of a fed-batch cultivation. Three potential model structures with less parameters than the original model were found after the calibration of each dataset (Table 14).

Table 14 – Potential robust structures for a fed-batch model

Structure	Parameters included
Original	$v_{MAX},\ K_S, v_{EtOH,B}, v_{Pyr,B}, v_{Arab,B}, v_{Cit,B},$ $v_{EtOH,FB}, v_{Pyr,FB}, v_{Arab,FB}, v_{Cit,FB}, \alpha_B, \alpha_{FB}, m_{ATP}, T_{Cons}$
1	$v_{MAX},\ K_S, v_{Pyr,B}, v_{Cit,B}, v_{EtOH,FB}, v_{Pyr,FB}, v_{Arab,FB}, v_{Cit,FB}, \alpha_B, m_{ATP}, T_{Cons}$
2	$K_S, v_{EtOH,B}, v_{Pyr,B}, v_{Arab,B}, v_{Cit,B}, v_{EtOH,FB}, v_{Pyr,FB}, \alpha_B, m_{ATP}$
3	$v_{MAX}, K_S, v_{Pyr,B}, v_{Arab,B}, v_{Cit,B}, v_{Pyr,FB}\ \alpha_B, \alpha_{FB}, m_{ATP}, T_{Cons}$

All the candidate structures consider the following parameters: K_S, $v_{PYR,B}$, $v_{CIT,B}$, α_B, $v_{PYR,FB}$ and m_{ATP}. Contrary to batch model, K_S. appears to play an important role for fed-batch cultivation. This parameter, which usually lies in the micromolar range (Boles & Hollenberg, 1997), can directly modulate substrate uptake under glucose-limited conditions. Therefore, when glucose concentration is close to zero (like in the feeding phase), slight variations in the value of K_S can change glucose uptake significantly, which has a direct impact in the specific growth rate. Also, m_{ATP} appears to have a relevant role as an energy sink when the glucose from the batch phase is depleted. Here, secondary product consumption starts and slower or null biomass formation rate is observed prior to the addition of glucose (around 30 hours of fermentation in Figure 8). This indicates that the substrates were consumed to maintain basic cellular functions to survive instead of being used for cell division.

The three *a priori* robust structures improved the initial fittings and reduced the frequency of parametric problems observed for the initial model of 14 parameters (Table 15). Structure 3 showed the best fitting capability compared to the original model. On average, this structure improved by 25% the initial calibrations. It is worthy to note that, despite Structure 3 did not include the minimum production rate of ethanol during the batch phase, it could adequately reproduce the profiles of this compound by calibrating the objective function and the maintenance ATP. Finally, Structure 3 was chosen to calibrate new fed-batch data in order to validate it as a robust modeling structure for glucose-limited aerobic fed-batch cultivations of *Pichia pastoris*.

Table 15 - Summary of the Cross Calibration of the fed-batch datasets. Structure that reduced de frequency of parametric problems with respect to the original model are highlighted.

Structure	N° parameters	F_{DIFF}	Significance Issues (%)	Sensitivity Issues (%)	Identifiability Issues (%)
Original	**14**	**0**	**33**	**18.8**	**3.9**
1	11	-2%	27.2	2.6	1.8
2	9	-15%	25.1	7.4	0.9
3	10	-25%	26.7	3.7	0.9

6.2.4. Robustness check of the chosen reduced modeling structure for glucose-limited aerobic fed-batch cultures of *Pichia pastoris*

Figure 9 illustrates the fit of Structure 3 to new experimental fed-batch data and Table 16 presents the parameter values with 95% confidence intervals achieved in the calibration.

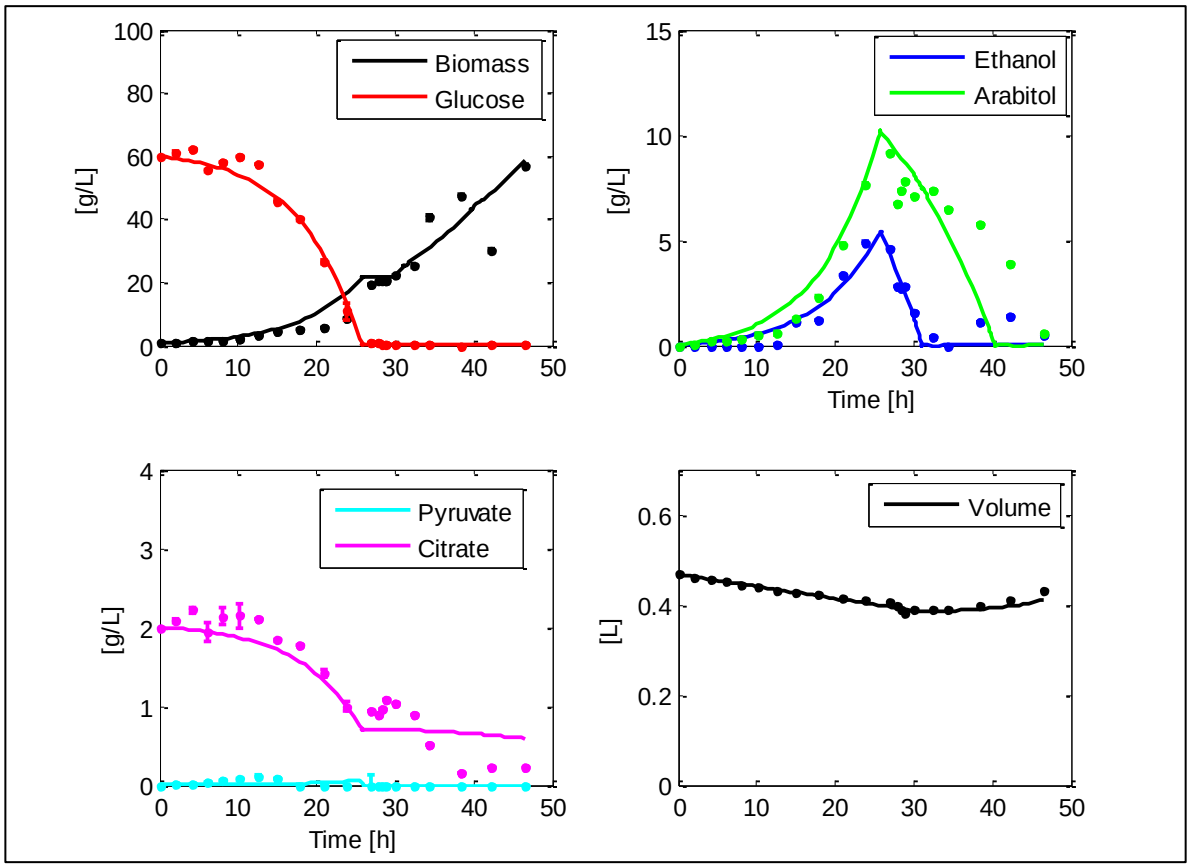


Figure 9 – Robustness check of Structure 3 as a robust representation of aerobic glucose-limited fed-batch cultures of *Pichia pastoris*

This calibration yielded no sensitivity problems. Also, parameters associated to pyruvate dynamics ($v_{PYR,B}$ and $v_{PYR,FB}$) presented significance problems due to the measurement noise at the near-zero concentrations found in this particular dataset. Finally, only one

identifiability issue between sub-optimal growth rate at the feed phase and the pyruvate consumption rate in this stage (Supplementary Material 8).

Pyruvate was found at such low concentrations that it might disregarded from the analysis by fixing $v_{PYR,B}$ and $v_{PYR,FB}$ at 0 mmol/g_{DCW}h. The calibration of the of this dataset considering no pyruvate formation, did not produce identifiability, significance or sensitivity problems.

Table 16 - Parameter values achieved in the calibration of the robustness check dataset using the fed-batch model. The confidence interval on the time where the consumption of secondary metabolites started T_{CONS}, could not be determined due to the stiffness of the solution caused by a sudden consumption of arabitol and ethanol.

Parameter	Value	Units
v_{MAX}	2.09 ± 0.46	mmol/g_{DCW}·h
K_S	$5.55 \cdot 10^{-2} \pm 0.0000004 \cdot 10^{-2}$	g/L
$v_{Pyr,B}$	0	mmol/g_{DCW}·h
$v_{Arab,B}$	0.42 ± 0.17	mmol/g_{DCW}·h
$v_{Cit,B}$	0.04 ± 0.00	mmol/g_{DCW}·h
$v_{Pyr,FB}$	0	mmol/g_{DCW}·h
α_B	$2.6 \cdot 10^{-4} \pm 0.4 \cdot 10^{-4}$	[-]
α_{FB}	$2.455 \cdot 10^{-5} \pm 0.003 \cdot 10^{-5}$	[-]
m_{ATP}	7.0 ± 1.4	mmol/g_{DCW}·h
T_{Cons}	25.73	h

The profile of some of the state variables still depended on the fixed values assigned to them. For example, in this particular experiment, arabitol was consumed at a slower rate than in the training dataset. The parameter representing this consumption was not included in the robust modeling structure and its value was fixed as the mean of the initial three datasets. The latter highlights the need to include more datasets from fed-batch cultivations, allowing the model to represent a wider variety of cellular dynamics.

However, the model structure showed a strong fitting capacity and a limited occurrence of identifiability, sensitivity and significance problems between parameters. Therefore,

we selected Structure 3 as the most robust modeling framework for fed-batch cultivations of *Pichia pastoris* found in our setup.

6.2.5. Fed-batch model validation

The parameters found in the best calibration of training set using the aforementioned robust (reduced) modeling structure yielded a good prediction of fed-batch dynamics. Biomass, glucose, ethanol and arabitol showed a good agreement with experimental data, whereas pyruvate and citrate dynamics were inaccurate. These differences appeared because in the culture from where the parameters were derived (Fed-batch culture 1, see Supplementary material 6) pyruvate formation was observed during the batch phase and citrate was formed during the feed phase, therefore the model assumed that these compounds were generated. Nevertheless, for the main most abundant compounds the model yielded a good prediction.

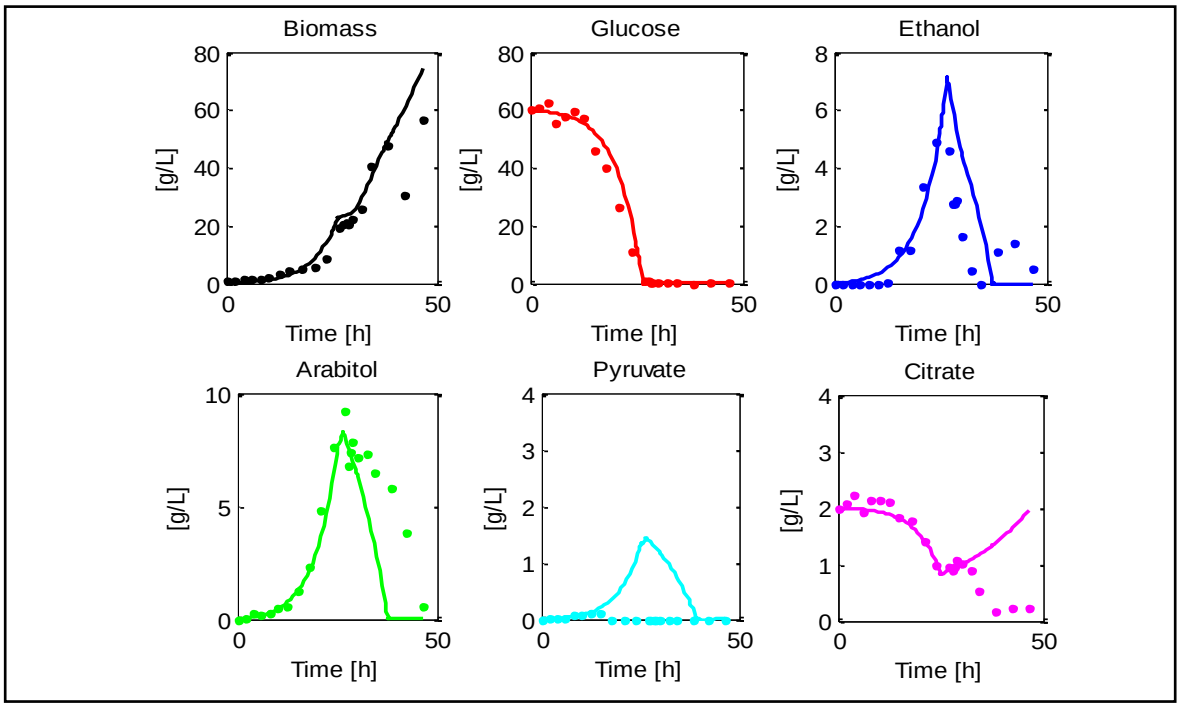


Figure 10 - Fed-batch model validation. Data is shown in filled dots while model prediction is presented in continuous lines.

6.3. Applications of the model

6.3.1. Model manual curation and analysis of the metabolic flux distribution during different stages of a dynamic cultivation

Once the model was validated and able to fit different experimental data without – or just a few – parametric problems, we checked the flux distributions that could be obtained during a fed-batch cultivation. Therefore, we chose three stages during the fermentation that represented different metabolic states of the cell (Figure 11):

1. Exponential growth and nutrient excess during the batch stage (~20 hrs, $\mu = 0.12\ h^{-1}$).
2. Co–consumption of arabitol and ethanol during glucose starvation (~27.5 hrs, $\mu = 0.02\ h^{-1}$).
3. Controlled exponential growth during the feeding stage (~45 hrs, $\mu = 0.06\ h^{-1}$).

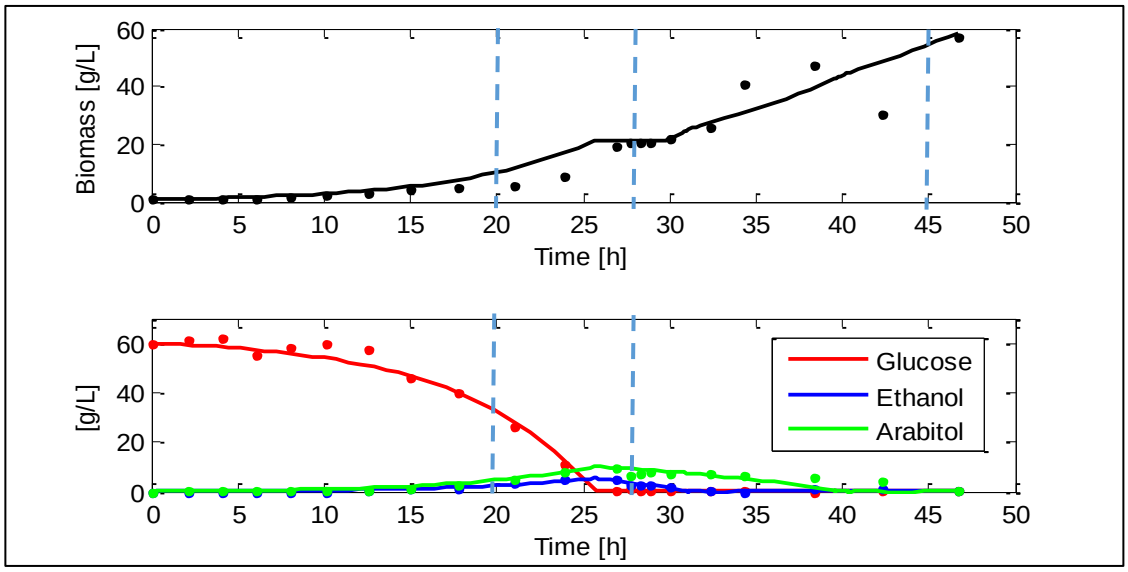


Figure 11. Biomass, glucose, ethanol and arabitol evolution of a fed-batch culture of *Pichia pastoris*. Dashed lines indicate the three instants of the fermentation where the metabolic flux distribution was determined.

Initially, the model did not carry flux through the oxidative part of the Pentose Phosphate Pathway (PPP), the main source of the reducing cofactor NADPH. Instead, this cofactor

was synthesized by a cytosolic NADP-dependent isocitrate dehydrogenase, which was also the main source of α-ketoglutarate in the cytosol (data not shown). Fluxomic studies in aerobic glucose-limited conditions in *P. pastoris* (Baumann et al., 2010; Dragosits et al., 2009; Heyland et al., 2010), have shown that about 40% of the carbon that reaches glucose 6 phosphate is carried through the oxidative branch of the PPP, which is also thermodynamically favorable (Nelson & Cox, 2008). Moreover, α-ketoglutarate is considered to be produced in the mitochondria and then exported to the cytosol for nitrogen fixation and anabolic reactions.

These inconsistencies have been recently addressed for several genome scale metabolic models of *Saccharomyces cerevisiae* (Pereira, Nielsen, & Rocha, 2016). Therefore, we performed the following changes to our reconstruction according to the indication of the authors:

i. We enabled the transport of α-ketoglutarate from the mitochondria to the cytosol using transporters present in *P. pastoris* (Rußmayer et al., 2015; Tomàs-Gamisans et al., 2016).
ii. The flux through three symporters that passively carried protons against the electrochemical gradient in the mitochondria was blocked in the direction of export to the cytosol.
iii. Based on the assumption postulated by Satrustegui *et al* (Satrustegui, Bautista, & Machado, 1983) for *Saccharomyces cerevisiae*, we considered that the NAD^+/NADH and NADPH/$NADP^+$ ratios in aerobic glucose-limited conditions are high enough to block the flux towards the formation of NAD^+ and NADPH. Therefore, we blocked 30 cytosolic reactions in the direction of either NAD^+ or NADPH formation. The only reactions left to produce cytosolic NADPH were the ones from the PPP and the acetate-forming acetaldehyde dehydrogenase, whose presence has been experimentally determined in aerobic glucose-limited cultivations of *P. pastoris* (Heyland et al., 2011).

Applying these modifications resulted in a spontaneous flux through the PPP, a mitochondrial formation of α-ketoglutarate with its subsequent secretion to the cytosol

and an overall concordance in the direction of the fluxes with respect to experimental data. This makes the model a reasonable approximation of *Pichia pastoris* central carbon metabolism.

When the flux distribution derived from the fed-batch robustness check dataset was compared with the fluxomic data obtained by Heyland *et al* (Heyland et al., 2010) (equivalent conditions), the average error in the prediction of 23 fluxes of the central metabolism dropped three times with respect to the predictions made by the not curated model - from 128% to 39% for the exponential batch phase and from 160% to 63% for the controlled feed phase (Figure 12). This drop was mainly caused by the change in the direction (from negative to positive flux) of the non-oxidative part of the PPP, the spontaneous flux through the oxidative branch of this pathway and the reduction in the predicted influx of oxaloacetate to the cell. The overall agreement in directionality can be seen in Figure 12 by the elimination of negative predicted fluxes in the curated model.

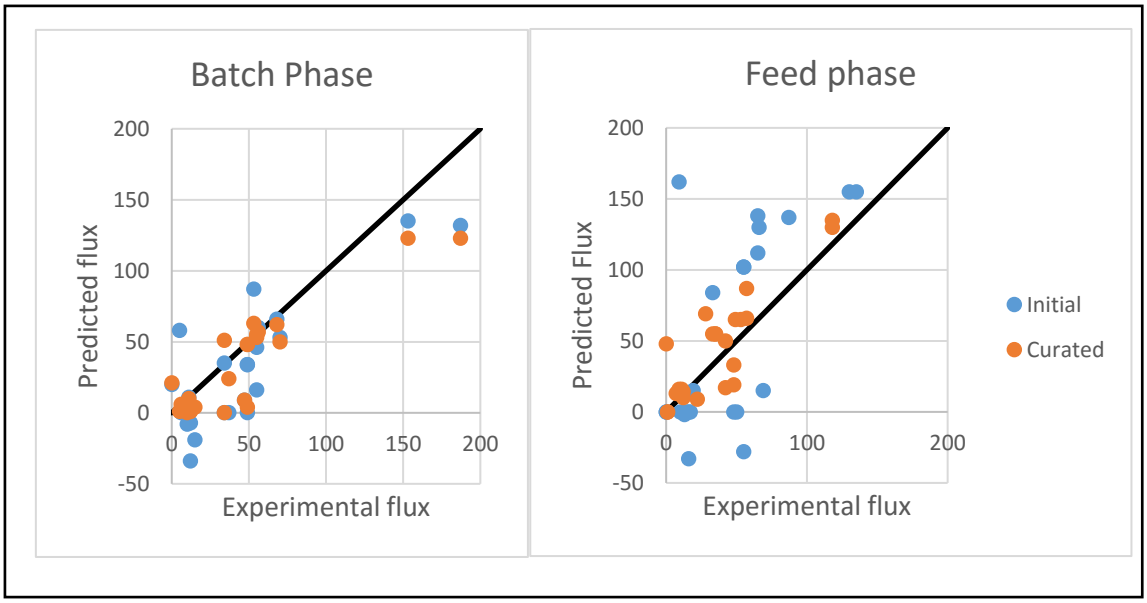


Figure 12 - Predicted versus experimental fluxes of the central metabolism. The flux distributions determined by Heyland *et al* (Heyland et al., 2010) during an aerobic glucose limited fermentation were compared to the output of the model in equivalent stages of a cultivation (exponential and controlled growth phases). Values are presented as normalized to carbon uptake and the black line represents the unit.

The flux distribution of the central carbon metabolism achieved with the curated model at the three fermentation times is presented in Figure 13. It is possible to distinguish several metabolic features in these stages:

1. Batch phase:
    - The carbon arriving to the glucose-6-phosphate node is split between carbohydrate production (11%), glycolysis (63%) and the oxidative branch of the PPP (24%). Despite carbon is divided between these pathways, flux through the PPP underestimates experimental values (40% of the incoming carbon).
    - Cytosolic NADPH is synthesized in the oxidative branch of the PPP.
    - Cytosolic ATP is formed by the activity of the ATP synthase and substrate-level phosphorylation from glycolysis as well as from the synthesis of arabitol and ethanol. ATP is mostly used for cell maintenance and biomass formation.

2. Starvation:
    - Ethanol and arabitol are co-consumed with limited formation of biomass (μ=0.02 h^{-1}) at this stage. Both compounds are directed towards the TCA cycle in order to form the reducing equivalents necessary to fuel the oxidative phosphorylation. The ATP formed in this pathway - ~ 7 mmol/g_{DCW}h -, is mostly employed for maintenance. Even though this m_{ATP} is high compared to other reported values for *P. pastoris* (2.2 – 5 mmol/g_{DCW}h), it is necessary to reproduce the observed limited cellular growth despite the fast consumption of secondary metabolites.

3. Controlled growth during the feed phase:
    - In this phase, neither ethanol nor arabitol were produced, and therefore all the carbon was directed towards biomass formation and the energy necessary for its synthesis and maintenance.

- In agreement previous fluxomic studies in aerobic glucose-limited chemostats (Baumann et al., 2010; Dragosits et al., 2009), flux through the oxidative and non-oxidative branches of the PPP without secretion of arabitol was observed. Furthermore, the model showed a high oxaloacetate transport from the cytosol to the mitochondria, which was also observed in the cited studies.
- The most distinguishable feature of this phase was the high activity of the TCA cycle, which almost doubled the flux through this pathway reported under glucose limited conditions ((Baumann et al., 2010; Dragosits et al., 2009; Heyland et al., 2011)). This is probably caused by the fact that our model considers the energy needed for maintenance, while more reduced reconstructions employed to derive TCA fluxes do not include this cost. Therefore, more activity in the TCA is needed to cope with maintenance and growth associated energy requirements, especially when no significant substrate level phosphorylation besides glycolysis is occurring.

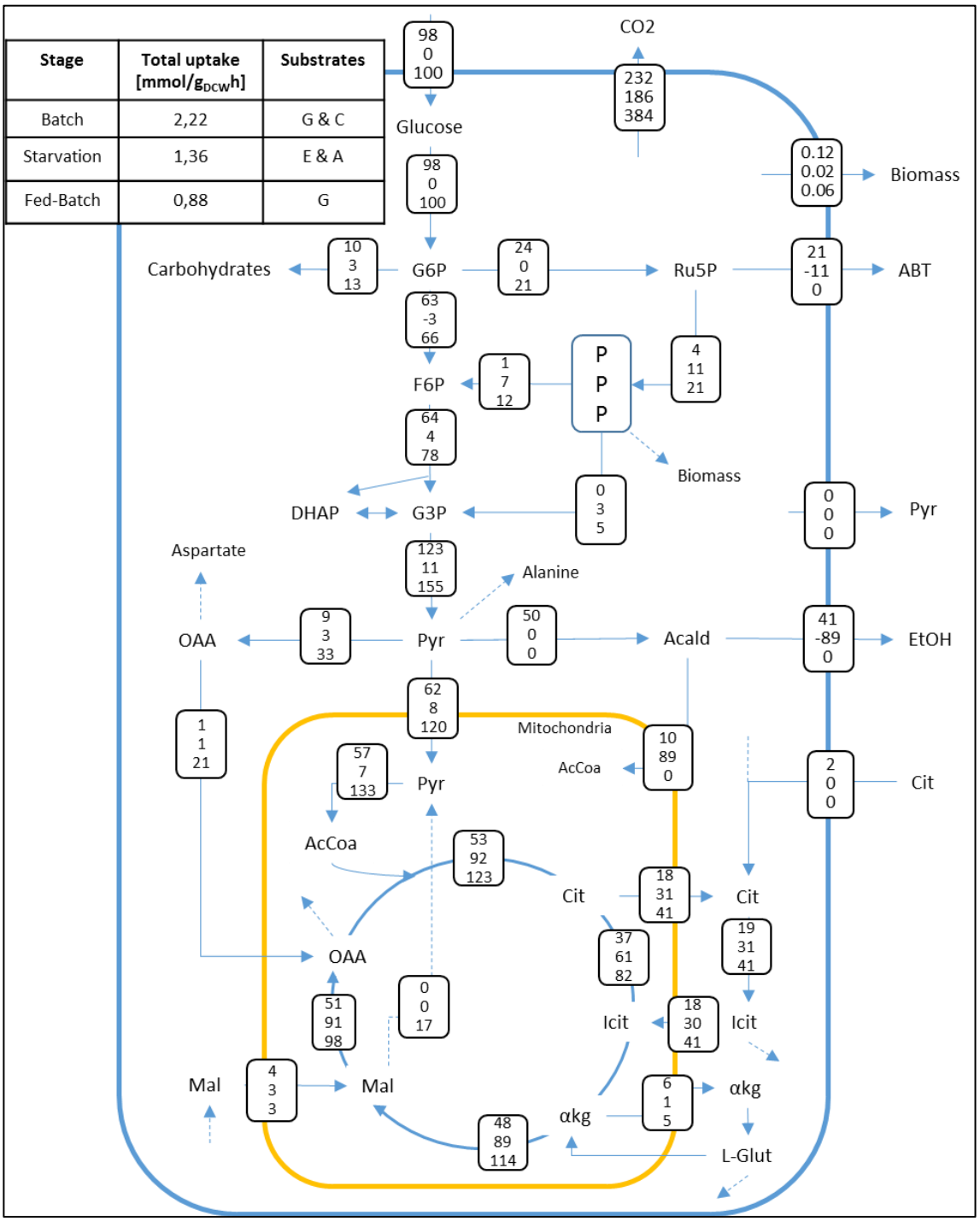


Figure 13 – Metabolic flux distribution in the Central metabolism for three different stages of the cultivation. Carbon uptake is detailed in the box of the upper left corner in mmol/g_{DCW}h and the fluxes are presented relative to this uptake. In each box between metabolites there are three numbers which correspond, from top to bottom, to the relative flux during batch, starvation and feeding phases. Depending on the time analyzed, the cell consumes Glucose (G), Citrate (C), Arabitol (A) or Ethanol (E). The biomass flux corresponds to the specific growth rate of the cell in h^{-1} and the negative fluxes refer to a change in the reaction directionality. Nomenclature for this figure is detailed in the Supplementary material 11.

To sum up, the model allowed the quantitative assessment of different physiological states present in aerobic fed-batch cultivations of *Pichia pastoris*. After model curation, the resulting flux distributions were reasonable in terms of reaction activity and directionality, but further curation is needed to improve the precision of the prediction, especially in the pentose phosphate pathway and the TCA cycle in glucose-limited conditions.

6.3.2. Discovery of single knock-outs to improve recombinant Human Serum Albumin production using Minimization of Metabolic Adjustment (MOMA) as objective function to simulate mutant behavior.

To demonstrate potential applications of the curated model, we performed 670 (number of genes in the model) batch simulations of single knock-out strains to discover beneficial deletions for the production of recombinant Human Serum Albumin (HSA), a 66 kDa protein with 16 disulfide bridges. HSA is a soluble, monomeric peptide that comprises about one half of the blood serum protein. It primarily acts as carrier of hormones and fatty acids and regulates the osmotic pressure of human blood(Verney, 1926).

The relation between final HSA and final biomass concentration of the 133 mutations that improved HSA production (> 30 mg/L at the end of the batch) are grouped into two main clusters (Figure 14). The first cluster consists in strains that privilege HSA production over biomass formation, whereas the second cluster present a trade-off between both.

We decided to leave Cluster I out of the analysis because of the impaired growth observed in the simulations, which was mainly due to the deletion of reactions associated to lipid biosynthesis. However, candidates from cluster II (32 in total) were manually analyzed to identify the cause of HSA overproduction (See Supplementary Material 10 for details on the genes).

A relative increase in the formation of cysteine and tryptophan was found for most of the candidates of Cluster II when compared to the parental strain, a trend that was not observed for the rest of the amino acids (Figure 15). These energetically costly residues

(Raiford et al., 2008) are formed from serine, therefore, re-routing carbon through this pathway could be beneficial to improve HSA production.

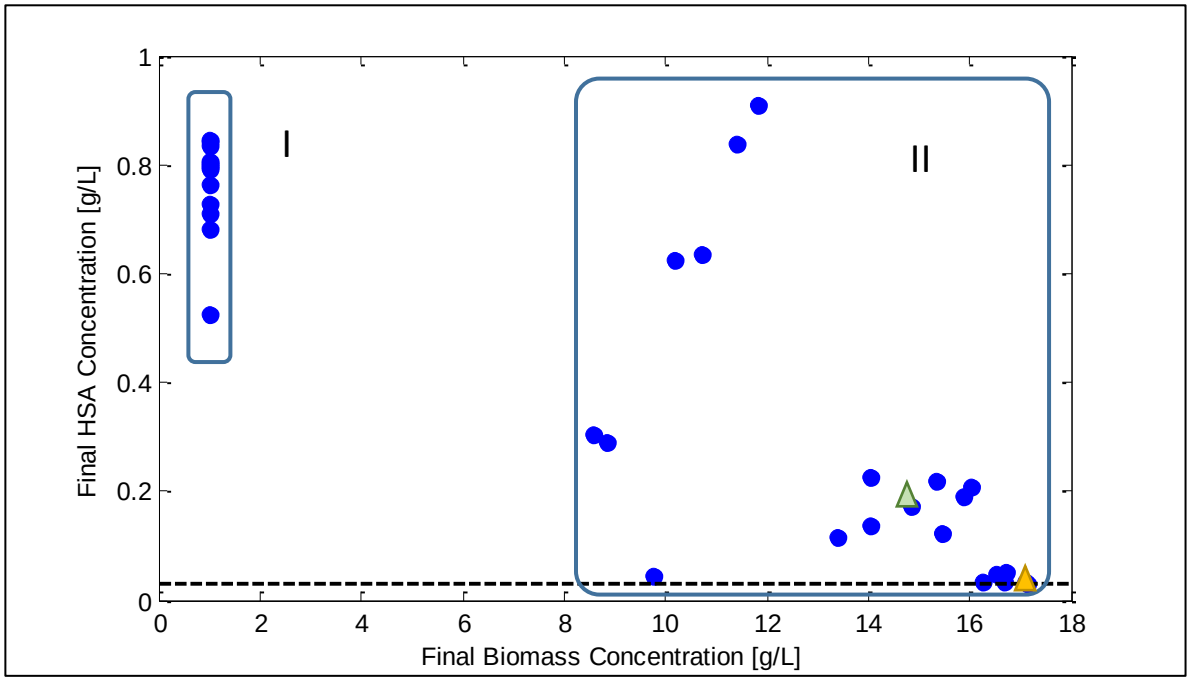


Figure 14 - Final HSA vs. final biomass concentrations of simulated batch cultivations of single knockout-strains. Blue dots correspond to the output of strains that improved the initial final HSA concentration (30 mg/L). Candidates out of Cluster II were manually analyzed. The yellow and green triangles indicate the performance of the Parental Strain and the methylene tetrahydrofolate dehydrogenase, respectively.

After analyzing the candidates, one strategy could be the deletion of the cytosolic NAD-dependent methylene tetrahydrofolate dehydrogenase. When compared to the parental strain, the knockout strain improved in 6.3 times the final concentration of recombinant protein (from 30 to 189 mg/L, triangles in Figure 14) with a 5.8-fold increase in protein volumetric productivity (from 2 to 11.5 mg/Lh). This deletion eliminates the drain of serine towards 5,10 methylene tetrahydrofolate, and serine can then be re-routed in two reactions to cysteine. This gene is non-essential in *Saccharomyces cerevisiae* (West, Horne, & Appling, 1996) and, to the best of our knowledge, its essentiality has not been

determined in *P. pastoris.* Therefore, it constitutes an interesting knockout candidate to improve recombinant HSA production.

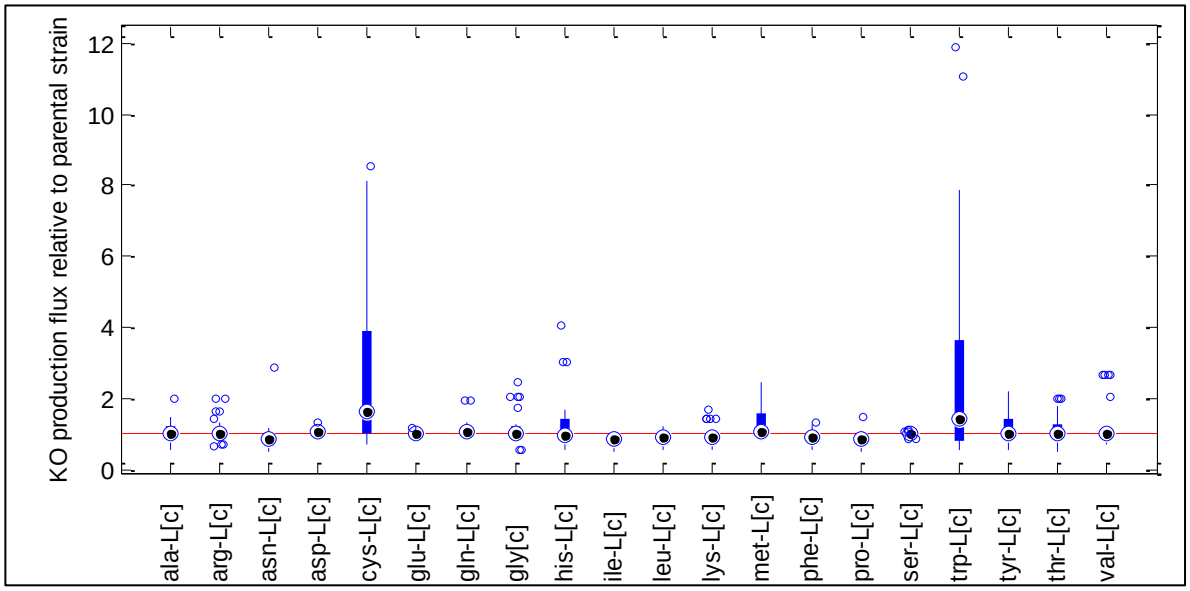


Figure 15 – Turnover of key amino acids in knock-out strains relative to the parental strain. Each box summarizes how the production of each amino acid changed in the 32 knock out strains of Cluster II relative to the production in the parental strain (Red Line). Black dots correspond to the sample median, the extreme of the boxes to the 25th and 75th percentiles, the whiskers extend to the most extreme data points and circles mark outliers.

6.3.3. Bioprocess optimization for the overproduction of Human Serum Albumin

The dynamic modeling framework could also be used for bioprocess optimization since it has the possibility to simulate cell behavior using information regarding strain physiology and process limitations (). In this exercise, we evaluated 13 feeding strategies of a fed-batch cultivation to improve the production of recombinant Human Serum Albumin (HSA) using the parental strain of the previous simulation as model.

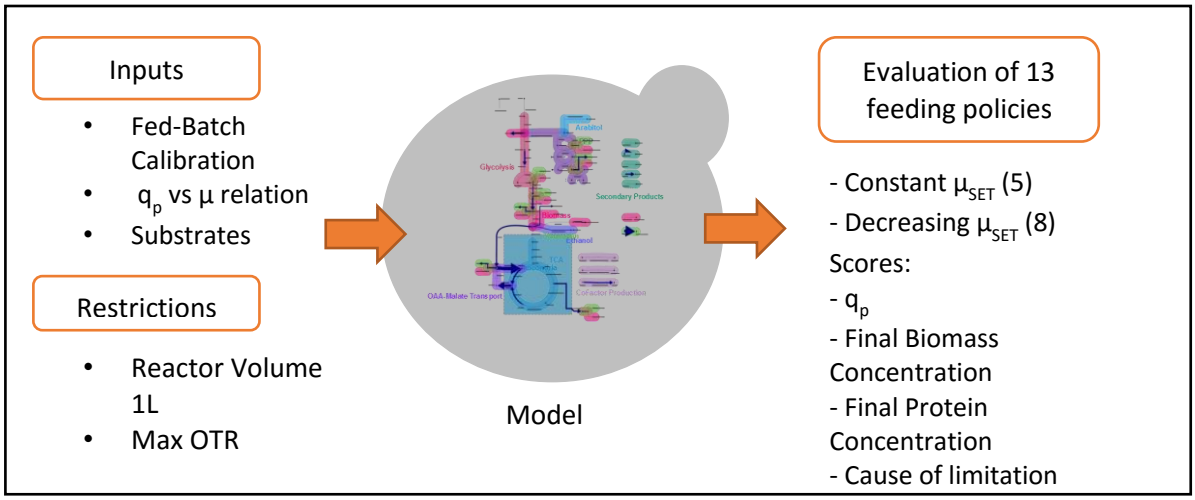


Figure 16 - Scheme of the bioprocess optimization problem.

After the simulations, we selected a strategy that considered a slow decrease in the growth rate from μ=0.14 h^{-1} to μ=0.08 h^{-1} during the feeding phase. When compared to cell growth at constant growth rates (Table 17), the chosen policy reached almost the same final HSA concentration as the constant growth rate strategy that reached the highest concentration (μ=0.06 h^{-1}), but with a 24% increase in volumetric productivity. This decreasing strategy managed to run the cultivation under the maximum oxygen transfer threshold while growing at high rates. The process was finally limited by reactor volume.

Table 17 - Feeding policies evaluated to improve the production of Serum Albumin in a particular bioreactor setup. This table shows the process indicators for the constant feeding Strategies (1-5) plus the best decreasing growth rate strategy. X_{FINAL} and P_{FINAL} refer to the final concentration of biomass and serum albumin in the reactor when the simulation stops, which happened by either violating user-defined volume and Oxygen Transfer thresholds.

Strategy	μ_{MAX}	Rate	μ_{MIN}	q_P [mg/gh]	X_{FINAL} [g/L]	P_{FINAL} [mg/L]	Limitation
1	0,14	-	-	2,85	164,8	138	Oxygen
2	0,12	-	-	2,59	187,8	135	Oxygen
3	0,1	-	-	2,32	195,3	130	Volume
4	0,08	-	-	2,29	191,3	138	Volume
5	0,06	-	-	2,28	184,7	154	Volume
Best	0,14	0,1	0,08	2,83	197,5	150	Volume

The improvement in process productivity by modifying substrate addition during the feed phase is much lower than the one attained by genetic modifications. However, other process variables such as reactor volume and oxygen transfer may be modified to further improve HSA production.

This example shows how the developed dFBA platform can be used as framework to organize information of a particular cultivation, to gain further insights of the strain at hand and to improve bioreactor performance.

7. CONCLUSIONS AND PERSPECTIVES

Current genome-scale metabolic models of *P. pastoris* have been employed to address cellular behavior in stationary conditions, successfully predicting production and consumption rates of different compounds and even improving by 40% recombinant protein production by model-discovered knock-outs (Nocon et al., 2014). However, little attention has been given to the actual metabolic flux distribution that these reconstructions yield and how they evolve in a dynamic environment. Resulting flux distributions are important for two reasons: (i) they help to understand the cellular response to the different stresses the cell is subjected to in this type of cultivation and (ii) they can serve as input for several algorithms whose aim is to find metabolic engineering targets to improve the production of a certain compound.

In this work, we developed a dynamic genome-scale metabolic model of glucose-limited aerobic cultivations of *Pichia pastoris*. The assembled platform is capable of fitting several datasets with minimum significance, sensitivity and identifiability problems in its parameters. It is also capable of predicting cellular dynamics if trained properly, where special attention should be payed to arabitol and pyruvate production. Moreover, the model could be used to obtain reasonable metabolic flux distributions throughout dynamic cultivations, which can be analyzed and used to determine of metabolic engineering

strategies to improve the production of a target compound *in silico*. The developed platform could therefore be employed to determine genetic and process engineering strategies to improve recombinant protein production.

The main findings of this study were:

- Adequate calibration of batch cultivations requires the inclusion of the maximum glucose uptake rate, the production rates of the species considered in the model and the sub-optimal growth coefficient in the adjustable parameter set. This ensures a proper fit and a minimal occurrence of parametric problems.
- A close fit to experimental data of fed-batch cultures requires the inclusion of maintenance ATP and the glucose uptake activity constant K_S in the adjustable parameter set.
- Blocking cytosolic reactions in the direction of NAD^+ and NADPH formation greatly improved the prediction of experimental fluxes of the central metabolism of *P. pastoris*.
- The resulting model allowed to identify metabolic engineering targets to improve recombinant protein production. From this analysis, we believe that cysteine and tryptophan biosynthesis pathways are potential limiting steps for the production of recombinant HSA. Also, the deletion of NAD-dependent methylene tetrahydrofolate dehydrogenase could potentially improve HSA volumetric productivity by 630% in batch cultivations by increasing cysteine formation
- The model was able to evaluate feeding policies that allowed an improvement in the production of Human Serum Albumin based on bioreactor restrictions, strain dynamics and the relation between μ and Protein productivity. An exponential feed rate that slowly decreases the specific growth rate at which a fed-batch culture was operated could potentially increase the productivity of HSA production by 24%.
- Gene knockouts proved to be more effective in the improvement of recombinant protein production than the modification of process variables.

The aforementioned findings present similarities and differences with previous works. First, Sanchez et al (Sánchez, Pérez-Correa, et al., 2014) determined a similar model structure for batch cultivations of *Saccharomyces cerevisiae*, but the fed-batch model structure they derived considered half the parameters used in this study. Unlike Sanchez *et al* in *S. cerevisiae*, we found that maintenance ATP could be properly estimated in fed-batch cultivations once glucose from the batch phase has been depleted and ethanol and arabitol consumption occurred without biomass growth.

It is worthy to mention two limitations of our approach. First, this model does not consider the calibration of gaseous mass balances of CO_2 and O_2. It predicts their production and consumption rates using the genome-scale model instead. Despite the good predicting capability of the model for these compounds (Supplementary material 1), it is desirable to include them in future versions of the platform since CO_2 formation can adjust the flux through reactions that use or produce this compound and dissolved oxygen concentration has a strong impact on process performance and can be easily manipulated (Baumann et al., 2010). Finally, the model disregards post-translational modifications and treats recombinant proteins as peptide chains (primary structures). The inclusion of glycosylation pathways and disulfide bond formation will recreate more accurately the stresses caused by the tertiary structure of proteins in *P. pastoris* metabolism (Irani et al., 2015).

There is a growing interest in *Pichia pastoris* metabolism. In fact, a new genome-scale metabolic reconstruction was recently released in January of 2016 with updated glycosylation pathways and a manually curated oxidative phosphorylation (Tomàs-Gamisans et al., 2016). This new model, called iMT1026, predicts experimental data from chemostats better than the iFS670 model used in this study (Supplementary Material 11). Therefore, it would be desirable to include it in our platform. Nevertheless, and as all the previously reported models, the internal flux distribution yielded by the reconstruction was not evaluated, so it is recommended to look at the internal fluxes before its use and perform a manual curation based on the logic derived by Pereira *et al* (Pereira et al., 2016), until reaching a reasonable flux distribution.

Also, special attention should be given to the objective function used to harness the knowledge contained in the genome scale metabolic models of *P. pastoris*. Despite that biomass optimization has been validated to represent its central metabolism in glucose limited conditions (Morales et al., 2014), it falls short when simulating its growth in glucose-rich environments where ethanol and arabitol were produced aerobically. If oxygen uptake is unconstrained, the model is unable to produce these compounds because all the carbon is directed towards growth when biomass formation is used as objective function. To overcome this limitation, we used parameter estimation to force the production of secondary metabolites. This suggests that for the accurate prediction of *P. pastoris* behavior in nutrient-rich environments other elements might be added to this objective function or more restrictions - derived from omics studies from example - should be applied in the directionality and maximum flux of certain reactions.

In order to broaden its applications in other conditions relevant for *P. pastoris*, the model could be calibrated with other carbon sources and feeding strategies such as glycerol batch phase followed by methanol induction during the feed phase. Also, the model could be used to study perturbations such as oxygen limitation, which is a common problem in industrial *P. pastoris* cultivations. Finally, it would be desirable to calibrate the model with data from a strain capable of producing high concentrations of a recombinant protein to understand and quantify the metabolic burden caused by this production.

In summary, this work corresponds to the first dynamic genome scale metabolic model assembled for *Pichia pastoris*, which was complemented with manual curation to obtain reasonable metabolic flux distributions during industrially relevant conditions. The developed platform may be used to understand cellular stresses during dynamic cultivations, to potentially predict bioreactor dynamics, and to propose metabolic and process engineering strategies for the improvement of recombinant protein production.

BIBLIOGRAPHY

Asadollahi, M. a., Maury, J., Patil, K. R., Schalk, M., Clark, A., & Nielsen, J. (2009). Enhancing sesquiterpene production in Saccharomyces cerevisiae through in silico driven metabolic engineering. *Metabolic Engineering*, *11*(6), 328–334. doi:10.1016/j.ymben.2009.07.001

Balsa-Canto, E., Rodriguez-Fernandez, M., & Banga, J. R. (2007). Optimal design of dynamic experiments for improved estimation of kinetic parameters of thermal degradation. *Journal of Food Engineering*, *82*(2), 178–188. doi:10.1016/j.jfoodeng.2007.02.006

Baumann, K., Carnicer, M., Dragosits, M., Graf, A. B., Stadlmann, J., Jouhten, P., … Ferrer, P. (2010). A multi-level study of recombinant Pichia pastoris in different oxygen conditions. *BMC Systems Biology*, *4*(1), 141. doi:10.1186/1752-0509-4-141

Baumann, K., Maurer, M., Dragosits, M., Cos, O., Ferrer, P., & Mattanovich, D. (2008). Hypoxic fed-batch cultivation ofPichia pastoris increases specific and volumetric productivity of recombinant proteins. *Biotechnology and Bioengineering*, *100*(1), 177–183. doi:10.1002/bit.21763

BCC Research. (2014). Global Markets for Enzymes in Industrial Applications - Report Overview. Retrieved December 7, 2015, from http://www.bccresearch.com/market-research/biotechnology/enzymes-industrial-applications-bio030h.html

Becker, S. a, Feist, A. M., Mo, M. L., Hannum, G., Palsson, B. Ø., & Herrgard, M. J. (2007). Quantitative prediction of cellular metabolism with constraint-based models: the COBRA Toolbox. *Nature Protocols*, *2*(3), 727–38. doi:10.1038/nprot.2007.99

Boles, E., & Hollenberg, C. P. (1997). The molecular genetics of hexose transport in yeasts. *FEMS Microbiology Reviews*, *21*(1), 85–111. doi:10.1016/S0168-6445(97)00052-1

Bornstein, B. J., Keating, S. M., Jouraku, A., & Hucka, M. (2008). LibSBML: An API library for SBML. *Bioinformatics*, *24*(6), 880–881. doi:10.1093/bioinformatics/btn051

Çalık, P., Ata, Ö., Güneş, H., Massahi, A., Boy, E., Keskin, A., … Özdamar, T. H. (2015). Recombinant protein production in Pichia pastoris under glyceraldehyde-3-phosphate dehydrogenase promoter: From carbon source metabolism to bioreactor operation parameters. *Biochemical Engineering Journal*, *95*, 20–36. doi:10.1016/j.bej.2014.12.003

Cárcamo, M. (2013). *Producción de Proteínas Recombinantes en Cultivos Fed-Batch de Saccharomyces Cerevisiae y Escharichia Coli*. Pontificia Unversidad Católica de Chile.

Cárcamo, M., Saa, P. A., Torres, J., Torres, S., Mandujano, P., Correa, J. R. P., & Agosin, E. (2014). Effective dissolved oxygen control strategy for high-cell-density cultures. *IEEE Latin America Transactions*, *12*(3), 389–

394. doi:10.1109/TLA.2014.6827863

Carnicer, M., Baumann, K., Töplitz, I., Sánchez-, F., Mattanovich, D., Ferrer, P., & Albiol, J. (2009). Macromolecular and elemental composition analysis and extracellular metabolite balances of Pichia pastoris growing at different oxygen levels, *14*, 1–14. doi:10.1186/1475-2859-8-65

Caspeta, L., Shoaie, S., Agren, R., Nookaew, I., & Nielsen, J. (2012). Genome-scale metabolic reconstructions of Pichia stipitis and Pichia pastoris and in silico evaluation of their potentials. *BMC Systems Biology*, *6*(1), 24. doi:10.1186/1752-0509-6-24

Cereghino, J. L., & Cregg, J. M. (2000). Heterologous protein expression in the methylotrophic yeast Pichia pastoris. *FEMS Microbiology Reviews*, *24*(1), 45–66.

Cheng, H., Lv, J., Wang, H., Wang, B., Li, Z., & Deng, Z. (2014). Genetically engineered Pichia pastoris yeast for conversion of glucose to xylitol by a single-fermentation process. *Applied Microbiology and Biotechnology*, *98*(8), 3539–3552. doi:10.1007/s00253-013-5501-x

Chung, B. K. S., Selvarasu, S., Camattari, A., Ryu, J., Lee, H., Ahn, J., … Lee, D. (2010). Genome-scale metabolic reconstruction and in silico analysis of methylotrophic yeast Pichia pastoris for strain improvement. *Microbial Cell Factories*, *9*(50), 2–15.

Chung, B. K.-S., Lakshmanan, M., Klement, M., Ching, C. B., & Lee, D.-Y. (2013). Metabolic reconstruction and flux analysis of industrial Pichia yeasts. *Applied Microbiology and Biotechnology*, *97*(5), 1865–73. doi:10.1007/s00253-013-4702-7

Ciofalo, V., Barton, N., Kreps, J., Coats, I., & Shanahan, D. (2006). Safety evaluation of a lipase enzyme preparation, expressed in Pichia pastoris, intended for use in the degumming of edible vegetable oil. *Regulatory Toxicology and Pharmacology*, *45*(1), 1–8. doi:10.1016/j.yrtph.2006.02.001

Čiplys, E., Žitkus, E., Gold, L. I., Daubriac, J., Pavlides, S. C., Højrup, P., … Slibinskas, R. (2015). High-level secretion of native recombinant human calreticulin in yeast. *Microbial Cell Factories*, *14*(1), 165. doi:10.1186/s12934-015-0356-8

Corchero, J. L., Gasser, B., Resina, D., Smith, W., Parrilli, E., Vázquez, F., … Villaverde, A. (2013). Unconventional microbial systems for the cost-efficient production of high-quality protein therapeutics. *Biotechnology Advances*, *31*(2), 140–153. doi:10.1016/j.biotechadv.2012.09.001

Daly, R., & Hearn, M. T. W. (2005). Expression of heterologous proteins in Pichia pastoris: a useful experimental tool in protein engineering and production. *Journal of Molecular Recognition : JMR*, *18*(2), 119–38. doi:10.1002/jmr.687

De Schutter, K., Lin, Y.-C., Tiels, P., Van Hecke, A., Glinka, S., Weber-Lehmann, J., … Callewaert, N. (2009).

Genome sequence of the recombinant protein production host Pichia pastoris. *Nature Biotechnology*, *27*(6), 561–6. doi:10.1038/nbt.1544

Delic, M., Göngrich, R., Mattanovich, D., & Gasser, B. (2014). Engineering of protein folding and secretion-strategies to overcome bottlenecks for efficient production of recombinant proteins. *Antioxidants & Redox Signaling*, *21*(3), 414–37. doi:10.1089/ars.2014.5844

Delic, M., Valli, M., Graf, A. B., Pfeffer, M., Mattanovich, D., & Gasser, B. (2013). The secretory pathway: exploring yeast diversity. *FEMS Microbiology Reviews*, *37*(6), 872–914. doi:10.1111/1574-6976.12020

Dragosits, M., Stadlmann, J., Albiol, J., Baumann, K., Maurer, M., Gasser, B., … Cerdanyola, B. (2009). The Effect of Temperature on the Proteome of Recombinant Pichia pastoris research articles. *Journal of Proteome Research*, (8), 1380–1392.

Egea, J., & Balsa-Canto, E. (2009). Dynamic optimization of nonlinear processes with an enhanced scatter search method. *Industrial & Engineering Chemical Research*, *48*(9), 4388–4401. Retrieved from http://pubs.acs.org/doi/abs/10.1021/ie801717t

Feng, X., Xu, Y., Chen, Y., & Tang, Y. J. (2012). Integrating flux balance analysis into kinetic models to decipher the dynamic metabolism of shewanella oneidensis MR-1. *PLoS Computational Biology*, *8*(2). doi:10.1371/journal.pcbi.1002376

Ferrer-Miralles, N., Domingo-Espín, J., Corchero, J. L., Vázquez, E., & Villaverde, A. (2009). Microbial factories for recombinant pharmaceuticals. *Microbial Cell Factories*, *8*, 17. doi:10.1186/1475-2859-8-17

Gasser, B., Prielhofer, R., Marx, H., Maurer, M., Nocon, J., Steiger, M., … Mattanovich, D. (2013). Pichia pastoris : protein production host and model organism for biomedical research. *Future Microbiology*, *8*(2), 191–208.

Graf, A., Dragosits, M., Gasser, B., & Mattanovich, D. (2009). Yeast systems biotechnology for the production of heterologous proteins. *FEMS Yeast Research*, *9*(3), 335–348. doi:10.1111/j.1567-1364.2009.00507.x

Hao, H., Zak, D. E., Sauter, T., Schwaber, J., & Ogunnaike, B. a. (2006). Modeling the VPAC2-activated cAMP/PKA signaling pathway: from receptor to circadian clock gene induction. *Biophysical Journal*, *90*(5), 1560–1571. doi:10.1529/biophysj.105.065250

Hasslacher, M., Schall, M., Hayn, M., Bona, R., Rumbold, K., Lückl, J., … Schwab, H. (1997). High-Level Intracellular Expression of Hydroxynitrile Lyase from the Tropical Rubber TreeHevea brasiliensisin Microbial Hosts. *Protein Expression and Purification*, *11*(1), 61–71. doi:http://dx.doi.org/10.1006/prep.1997.0765

Heyland, J., Fu, J., Blank, L. M., & Schmid, A. (2010). Quantitative physiology of Pichia pastoris during glucose-limited high-cell density fed-batch cultivation for recombinant protein production. *Biotechnology and*

Bioengineering, *107*(2), 357–368. doi:10.1002/bit.22836

Heyland, J., Fu, J., Blank, L. M., & Schmid, A. (2011). Carbon metabolism limits recombinant protein production in Pichia pastoris. *Biotechnology and Bioengineering*, *108*(8), 1942–53. doi:10.1002/bit.23114

Holzhütter, H. G. (2004). The principle of flux minimization and its application to estimate stationary fluxes in metabolic networks. *European Journal of Biochemistry*, *271*(14), 2905–2922. doi:10.1111/j.1432-1033.2004.04213.x

Hyduke, D., Schellenberger, J., Que, R., Fleming, R., Thiele, I., Orth, J., … Palsson, B. (2011). COBRA Toolbox 2.0. *Protocol Exchange*. doi:10.1038/protex.2011.234

Irani, Z. A., Kerkhoven, E. J., Shojaosadati, S. A., & Nielsen, J. (2015). Genome-scale metabolic model of Pichia pastoris with native and humanized glycosylation of recombinant proteins. *Biotechnology and Bioengineering*, *113*(5), 961–969. doi:10.1002/bit.25863

Keating, S. M., Bornstein, B. J., Finney, A., & Hucka, M. (2006). SBMLToolbox: An SBML toolbox for MATLAB users. *Bioinformatics*, *22*(10), 1275–1277. doi:10.1093/bioinformatics/btl111

Kitano, H. (2002). Systems biology: a brief overview. *Science (New York, N.Y.)*, *295*(5560), 1662–4. doi:10.1126/science.1069492

Landaw, E. M., & DiStefano 3rd, J. J. (1984). Multiexponential, multicompartmental, and noncompartmental modeling. II. Data analysis and statistical considerations. *American Journal of Physiology - Regulatory, Integrative and Comparative Physiology*, *246*(5).

Looser, V., Bruhlmann, B., Bumbak, F., Stenger, C., Costa, M., Camattari, A., … Kovar, K. (2015). Cultivation strategies to enhance productivity of Pichia pastoris : A review, *33*, 1177–1193.

Maccani, A., Landes, N., Stadlmayr, G., Maresch, D., Leitner, C., Maurer, M., … Mattanovich, D. (2014). Pichia pastoris secretes recombinant proteins less efficiently than Chinese hamster ovary cells but allows higher space-time yields for less complex proteins. *Biotechnology Journal*, *9*(4), 526–37. doi:10.1002/biot.201300305

Mahadevan, R., Edwards, J. S., & Doyle, F. J. (2002). Dynamic flux balance analysis of diauxic growth in Escherichia coli. *Biophysical Journal*, *83*(3), 1331–1340. doi:10.1016/S0006-3495(02)73903-9

Markets and Markets. (2015). Industrial Enzymes Market by Type (Carbohydrases, Proteases, Non-starch Polysaccharides & Others), Application (Food & Beverage, Cleaning Agents, Animal Feed & Others), Brands & by Region - Globlar Trends and Forecasts to 2020. Retrieved December 7, 2015, from http://www.marketsandmarkets.com/Market-Reports/industrial-enzymes-market-237327836.html

Marx, H., Mecklenbräuker, A., Gasser, B., Sauer, M., & Mattanovich, D. (2009). Directed gene copy number amplification in Pichia pastoris by vector integration into the ribosomal DNA locus. *FEMS Yeast Research*,

9(8), 1260–1270. doi:10.1111/j.1567-1364.2009.00561.x

Masuda, T., Ide, N., Ohta, K., & Kitabatake, N. (2010). High-yield Secretion of the Recombinant Sweet-Tasting Protein Thaumatin I. *Food Science and Technology Research*, *16*(6), 585–592.

Mattanovich, D., Graf, A., Stadlmann, J., Dragosits, M., Redl, A., Maurer, M., … Gasser, B. (2009). Genome, secretome and glucose transport highlight unique features of the protein production host Pichia pastoris. *Microbial Cell Factories*, *8*, 29. doi:10.1186/1475-2859-8-29

Maurer, M., Kühleitner, M., Gasser, B., & Mattanovich, D. (2006). Versatile modeling and optimization of fed batch processes for the production of secreted heterologous proteins with Pichia pastoris. *Microbial Cell Factories*, *5*, 37. doi:10.1186/1475-2859-5-37

Morales, Y., Tortajada, M., Picó, J., Vehí, J., & Llaneras, F. (2014). Validation of an FBA model for Pichia pastoris in chemostat cultures. *BMC Systems Biology*, *8*(1), 142. doi:10.1186/s12918-014-0142-y

Nelson, D. L., & Cox, M. M. (2008). *LEHNINGER, PRINCIPLES OF BIOCHEMISTRY* (Fifth Edit.). New York: W. H. Freeman and company.

Nocon, J., Steiger, M. G., Pfeffer, M., Sohn, S. B., Kim, T. Y., Maurer, M., … Mattanovich, D. (2014). Model based engineering of Pichia pastoris central metabolism enhances recombinant protein production. *Metabolic Engineering*, *24*, 129–38. doi:10.1016/j.ymben.2014.05.011

Orth, J. D., Thiele, I., & Palsson, B. Ø. (2010). What is flux balance analysis? *Nature Biotechnology*, *28*(3), 245–8. doi:10.1038/nbt.1614

Overton, T. W. (2014). Recombinant protein production in bacterial hosts. *Drug Discovery Today*, *19*(5), 590–601. doi:10.1016/j.drudis.2013.11.008

Palsson, B. O. (2015). *Systems Biology: Constraint-based Reconstruction and Analysis*. Cambridge: Cambridge University Press.

Park, J. H., Lee, K. H., Kim, T. Y., & Lee, S. Y. (2007). Metabolic engineering of Escherichia coli for the production of L-valine based on transcriptome analysis and in silico gene knockout simulation. *Proceedings of the National Academy of Sciences of the United States of America*, *104*(19), 7797–7802. doi:10.1073/pnas.0702609104

Pereira, R., Nielsen, J., & Rocha, I. (2016). Improving the flux distributions simulated with genome-scale metabolic models of Saccharomyces cerevisiae. *Metabolic Engineering Communications*, *3*, 153–163. doi:10.1016/j.meteno.2016.05.002

Petersen, B., Gernaey, K., & Vanrolleghem, P. A. (2001). Practical identifiability of model parameters by combined respirometric-titrimetric measurements. *Water Science and Technology*, *43*(7), 347–355.

Postma, E., Verduyn, C., Scheffers, W. a., & Van Dijken, J. P. (1989). Enzymic analysis of the crabtree effect in glucose-limited chemostat cultures of Saccharomyces cerevisiae. *Applied and Environmental Microbiology*, *55*(2), 468–477.

Price, N. D., Famili, I., Beard, D. a, & Palsson, B. Ø. (2002). Extreme pathways and Kirchhoff's second law. *Biophysical Journal*, *83*(5), 2879–82. doi:10.1016/S0006-3495(02)75297-1

Prielhofer, R., Maurer, M., Klein, J., Wenger, J., Kiziak, C., Gasser, B., & Mattanovich, D. (2013). Induction without methanol: novel regulated promoters enable high-level expression in Pichia pastoris. *Microbial Cell Factories*, *12*(1), 5. doi:10.1186/1475-2859-12-5

Raiford, D. W., Heizer, E. M., Miller, R. V, Akashi, H., Raymer, M. L., & Krane, D. E. (2008). Do amino acid biosynthetic costs constrain protein evolution in Saccharomyces cerevisiae? *Journal of Molecular Evolution*, *67*(6), 621–30. doi:10.1007/s00239-008-9162-9

Rebnegger, C., Graf, A. B., Valli, M., Steiger, M. G., Gasser, B., Maurer, M., & Mattanovich, D. (2014). In Pichia pastoris, growth rate regulates protein synthesis and secretion, mating and stress response. *Biotechnology Journal*, *9*(4), 511–25. doi:10.1002/biot.201300334

Rußmayer, H., Buchetics, M., Gruber, C., Valli, M., Grillitsch, K., Modarres, G., … Gasser, B. (2015). Systems-level organization of yeast methylotrophic lifestyle. *BMC Biology*, *13*(1), 80. doi:10.1186/s12915-015-0186-5

Sacher, J., Saa, P., Cárcamo, M., López, J., Gelmi, C. a., & Pérez-Correa, R. (2011). Improved calibration of a solid substrate fermentation model. *Electronic Journal of Biotechnology*, *14*(5). doi:10.2225/vol14-issue5-fulltext-7

Sánchez, B. J., Pérez-Correa, J. R., & Agosin, E. (2014). Construction of robust dynamic genome-scale metabolic model structures of Saccharomyces cerevisiae through iterative re-parameterization. *Metabolic Engineering*, *25*, 159–73. doi:10.1016/j.ymben.2014.07.004

Sánchez, B. J., Soto, D. C., Jorquera, H., Gelmi, C. A., & Pérez-Correa, J. R. (2014). HIPPO: An iterative reparametrization method for identification and calibration of dynamic bioreactor models of complex processes. *Industrial and Engineering Chemistry Research*, *53*(48), 18514–18525. doi:10.1021/ie501298b

Satrustegui, J., Bautista, J., & Machado, A. (1983). NADPH/NADP+ Ratio - Regulatory Implications in Yeast Glyoxylic-Acid Cycle. *Molecular and Cellular Biochemistry*, *51*(2), 123–127.

Schuetz, R., Kuepfer, L., & Sauer, U. (2007). Systematic evaluation of objective functions for predicting intracellular fluxes in Escherichia coli. *Molecular Systems Biology*, *3*(119), 119. doi:10.1038/msb4100162

Schuetz, R., Zamboni, N., Zampieri, M., Heinemann, M., & Sauer, U. (2012). Multidimensional optimality of microbial metabolism. *Science (New York, N.Y.)*, *336*(6081), 601–4. doi:10.1126/science.1216882

Segrè, D., Vitkup, D., & Church, G. M. (2002). Analysis of optimality in natural and perturbed metabolic networks. *PNAS*, *99*(23), 15112–15117.

Sohn, S. B., Graf, A. B., Kim, T. Y., Gasser, B., Maurer, M., Ferrer, P., … Lee, S. Y. (2010). Genome-scale metabolic model of methylotrophic yeast Pichia pastoris and its use for in silico analysis of heterologous protein production. *Biotechnology Journal*, *5*(7), 705–15. doi:10.1002/biot.201000078

Solà, A., Jouhten, P., Maaheimo, H., Sánchez-Ferrando, F., Szyperski, T., & Ferrer, P. (2007). Metabolic flux profiling of Pichia pastoris grown on glycerol/methanol mixtures in chemostat cultures at low and high dilution rates. *Microbiology (Reading, England)*, *153*(Pt 1), 281–90. doi:10.1099/mic.0.29263-0

Sriram, K., Rodriguez-Fernandez, M., & Doyle, F. J. (2012). Modeling cortisol dynamics in the neuro-endocrine axis distinguishes normal, depression, and post-traumatic stress disorder (PTSD) in humans. *PLoS Computational Biology*, *8*(2). doi:10.1371/journal.pcbi.1002379

Stephanopoulos, G. M., Aristidou, A. A., & Nielsen, J. (1998). *Metabolic Engineering Principles and Methodologies*. San Diego: Academic Press.

Thiele, I., & Palsson, B. Ø. (2010). A protocol for generating a high-quality genome-scale metabolic reconstruction. *Nature Protocols*, *5*(1), 93–121. doi:10.1038/nprot.2009.203

Thompson, C. A. (2010). FDA approves kallikrein inhibitor to treat hereditary angioedema. *American Journal of Health-System Pharmacy*, *67*, 93. doi:10.2146/news100005

Tolner, B., Smith, L., Begent, R. H. J., & Chester, K. a. (2006). Production of recombinant protein in Pichia pastoris by fermentation. *Nature Protocols*, *1*(2), 1006–1021. doi:10.1038/nprot.2006.126

Tomàs-Gamisans, M., Ferrer, P., & Albiol, J. (2016). Integration and Validation of the Genome-Scale Metabolic Models of Pichia pastoris: A Comprehensive Update of Protein Glycosylation Pathways, Lipid and Energy Metabolism. *Plos One*, *11*(1), e0148031. doi:10.1371/journal.pone.0148031

Van Bodegom, P. (2007). Microbial maintenance: A critical review on its quantification. *Microbial Ecology*, *53*(4), 513–523. doi:10.1007/s00248-006-9049-5

van Urk, H., Postma, E., Scheffers, W. a, & van Dijken, J. P. (1989). Glucose transport in crabtree-positive and crabtree-negative yeasts. *Journal of General Microbiology*, *135*(1987), 2399–2406. doi:10.1099/00221287-135-9-2399

Varela, C. a, Baez, M. E., & Agosin, E. (2004). Osmotic Stress Response : Quantification of Cell Maintenance and Metabolic Fluxes in a Lysine-Overproducing Strain of Corynebacterium glutamicum Osmotic Stress Response : Quantification of Cell Maintenance and Metabolic Fluxes in a Lysine-Overproducing St, *70*(7), 4222–4229. doi:10.1128/AEM.70.7.4222

Varma, A., & Palsson, B. (1994). Stoichiometric Flux Balance Models Quantitatively Predict Growth and Metabolic By-Product Secretion in Wild-Type Escherichia coli W3110, *60*(10), 3724–3731.

Verney, E. B. (1926). The osmotic pressure of the proteins of human serum and plasma. *The Journal of Physiology*, *61*(3), 319–328.

Villadsen, J., Nielsen, J., & Lidén, G. (2011). *Bioreaction Engineering Principles* (3rd ed.). Nueva York: Springer.

Villadsen, J., & Patil, K. R. (2007). Optimal Fed-Batch Cultivation When Mass Transfer Becomes Limiting. *Biotechnology and Bioengineering*, *98*(3), 706–710.

Walsh, G. (2014). Biopharmaceutical benchmarks 2014. *Nature Biotechnology*, *32*(7), 992–1000. doi:10.1038/nbt0910-917

Wang, J.-R., Li, Y.-Y., Liu, D.-N., Liu, J.-S., Li, P., Chen, L.-Z., & Xu, S.-D. (2015). Codon Optimization Significantly Improves the Expression Level of α -Amylase Gene from *Bacillus licheniformis* in *Pichia pastoris*. *BioMed Research International*, *2015*, 1–9. doi:10.1155/2015/248680

Wang, Y., Liang, Z. H., Zhang, Y. S., Yao, S. Y., Xu, Y. G., Tang, Y. H., … Feng, Y. M. (2001). Human insulin from a precursor overexpressed in the methylotrophic yeast Pichia pastoris and a simple procedure for purifying the expression product. *Biotechnology and Bioengineering*, *73*, 74–79. doi:10.1002/1097-0290(20010405)73:1<74::AID-BIT1038>3.0.CO;2-V

West, M. G., Horne, D. W., & Appling, D. R. (1996). Metabolic role of cytoplasmic isozymes of 5,10-methylenetetrahydrofolate dehydrogenase in Saccharomyces cerevisiae. *Biochemistry*, *35*(9), 3122–3132. doi:10.1021/bi952713d

SUPPLEMENTARY MATERIAL

I. **Supplementary Material 1:** Modifications performed over the iPP618 model and genome-scale models comparison

Modifications performed on the iPP618 model

Addition of the D-Arabitol synthesis pathway

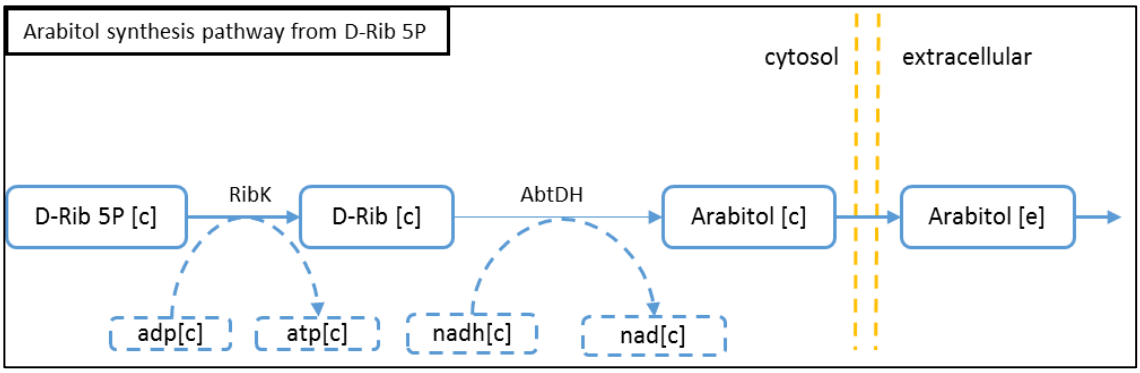


Figure 17 – D-Arabitol synthesis pathway from D-Ribulose-5-phosphate in Pichia pastoris. In total, we added four reactions associated to this pathway. First, Ribulose-5P is converted into D-Ribose by a kinase with the formation of ATP. Then, D-ribose is converted into D-arabitol by a dehydrogenase with the formation of nad from nadh. After D-arabitol is synthesized, it is transported to the extracellular medium and then "consumed" by an exchange reaction.

Addition of thaumatin, Human Serum Albumin (HSA) and FAB synthesis pathways

Thaumatin, HSA and FAB synthesis pathways were also included in the model according to the DNA, RNA and amino acid requirements employed in the iLC915 model to form the primary structure of the protein:

$$0.997 \cdot AA_{Thau} + 0.0029 \cdot RNA_{Thau} + 0.000028 \cdot DNA_{Thau} \rightarrow Thaumatin\ [c] \quad (23)$$

$$\left(\sum_{1}^{20} \beta_i \cdot aa_i\right) + \gamma \cdot ATP[c] + \gamma \cdot H_2O[c] \rightarrow AA_{Thau} \quad (24)$$

$$\left(\sum_{1}^{4} \alpha_i \cdot NMP_i\right) + \delta \cdot ATP[c] + \delta \cdot H_2O[c] \rightarrow RNA_{Thau} \quad (25)$$

$$\left(\sum_{1}^{4} \varepsilon_i \cdot dNMP_i\right) + \theta \cdot ATP[c] + \theta \cdot H_2O[c] \rightarrow RNA_{Thau} \quad (26)$$

Coefficients for the different components of the proteins are detailed in Tables 18, 19 and 20. Codon usage was taken from (De Schutter et al., 2009) and was used as input for the calculation of RNA and DNA sequences online (http://www.bioinformatics.org/sms2/rev_trans.html).

Table 18 - Amino acid requirements to form 1 gram of thaumatin, HSA and FAB fragment in the iPP669 model. The coefficients here reported were used in equation 2, a cost of 4,3 mole of ATP was assumed per mole of amino acid assembles in the protein. All coefficients have mmol/gram of protein units.

Substrate	Thaumatin	HSA	Fab Fragment
L-Alanine	0,721	0,909	0,559
L-Arginine	0,541	0,390	0,430
L-Asparagine	0,451	0,245	0,215
L-Aspartate	0,541	0,519	0,387
L-Cysteine	0,721	0,505	0,215
L-Glutamate	0,271	0,895	0,473
L-Glutamine	0,180	0,289	0,559
Glycine	1,082	0,188	0,602
L-Histidine	0,000	0,231	0,172
L-Isoleucine	0,361	0,130	0,215
L-Leucine	0,406	0,923	0,774
L-Lysine	0,496	0,866	0,387
L-Methionine	0,045	0,101	0,043
L-Phenylalanine	0,496	0,505	0,387
L-Proline	0,541	0,346	0,516
L-Serine	0,631	0,404	1,376
L-Threonine	0,902	0,418	0,731
L-Tryptophan	0,135	0,029	0,086
L-Tyrosine	0,361	0,274	0,387
L-Valine	0,451	0,620	0,645
ATP (γ)	40,1	37,8	39,4

Table 19 - RNA requirements for the production of 1 gram of Thaumatin, HSA and Fab fragment codifying RNA. A cost of 2,4 mol of ATP per gram of RNA was assumed.

Substrate	Thaumatin	HSA	Fab Fragment
AMP	0,73	0,86	0,74
UMP	1,14	1,09	1,18
GMP	0,73	0,72	0,58
CMP	0,48	0,40	0,57
ATP (δ)	7,38	7,38	7,38

Table 20 - DNA requirement for the formation of 1 gram of codifying DNA for Thaumatin, HSA and Fab Fragment. A cost of 3,4 mol ATP per gram of DNA produced was assumed.

Substrate	Thaumatin	HSA	Fab Fragment
dAMP	1,14	0,86	1,18
dTMP	0,73	1,09	0,74
dGMP	0,48	0,72	0,57
dCMP	0,73	0,40	0,58
ATP(θ)	10,45	10,45	10,45

Model Comparison

Figures 18, 19 and 20 show the performance of the different models to predict experimental growth rates.

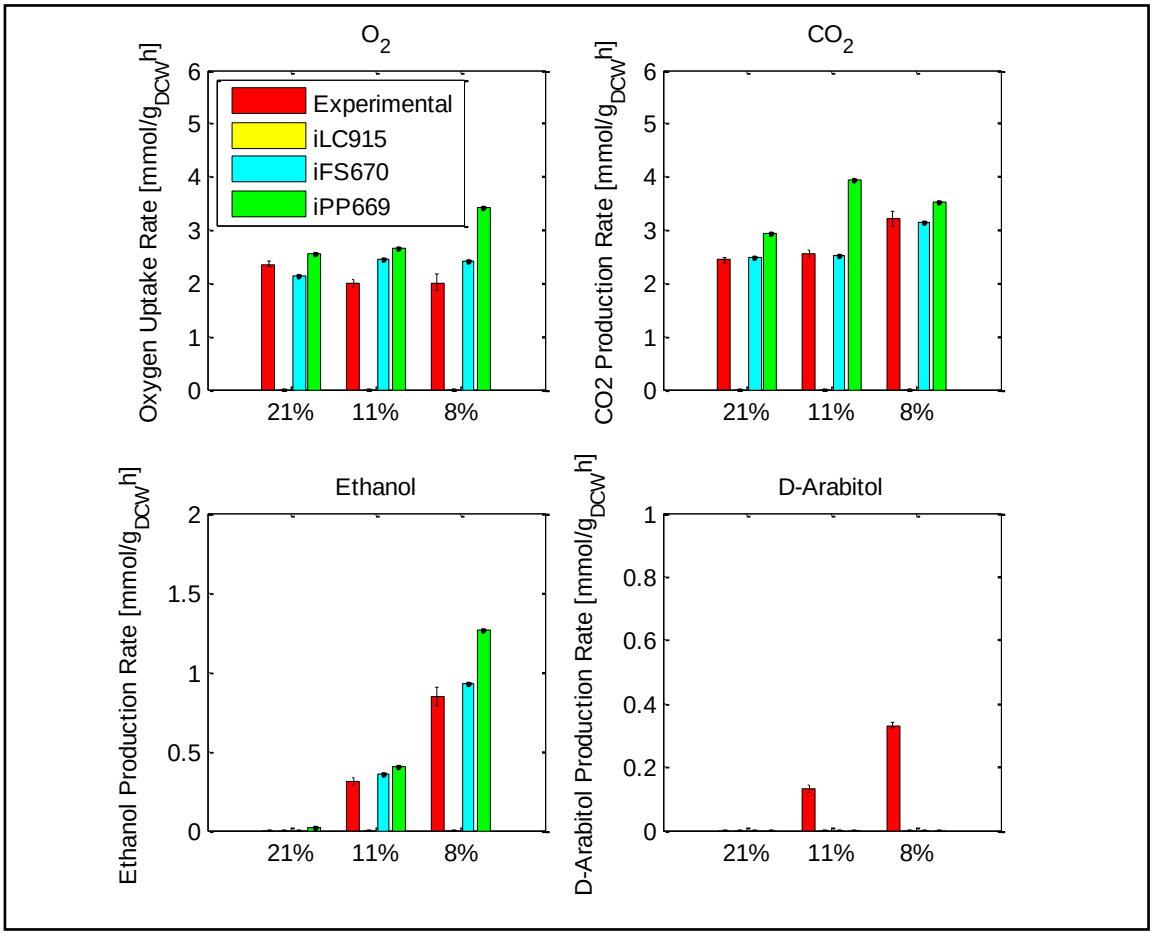


Figure 18 - Gas exchange and secondary metabolite prediction from Dataset 2 by the analyzed models. The predictions were performed with the uptake and production rates obtained for the parental strain, the same analysis was done with the recombinant strain yielding similar results. The percentage in the x axis correspond to oxygen fraction in the gas inlet of the bioreactor used to perform the study (21% → normoxic, 11% → oxygen limited, 8% → hypoxic)

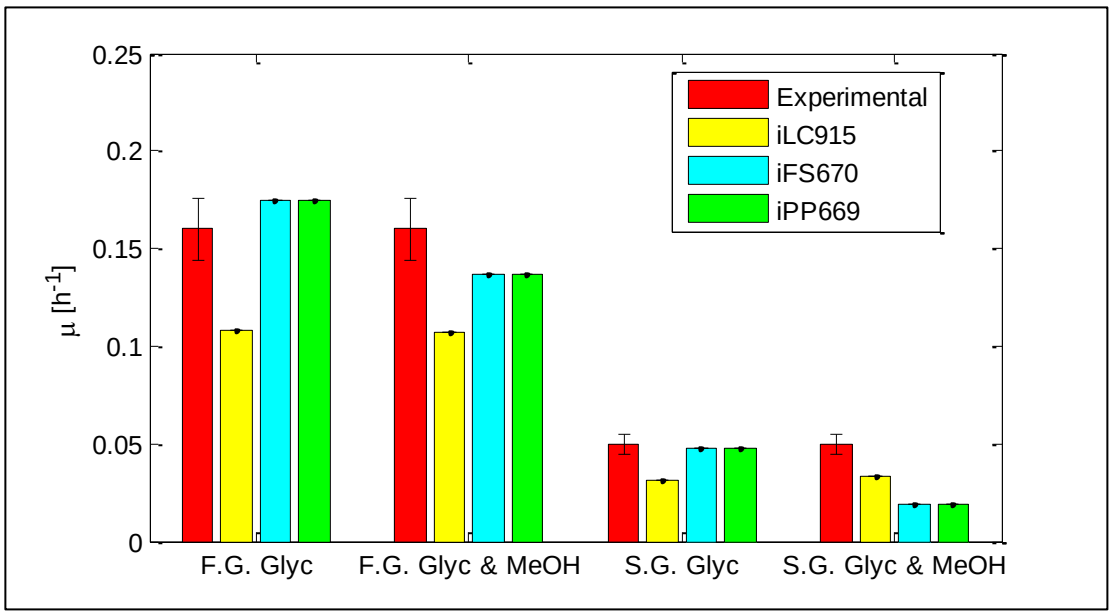


Figure 19 - Growth rate prediction in glycerol and glycerol/methanol limited chemostats. F.G. stands for Fast Growth (μ=0.16 h^{-1}) and S.G. means Slow Growth (μ=0.05 h^{-1}).

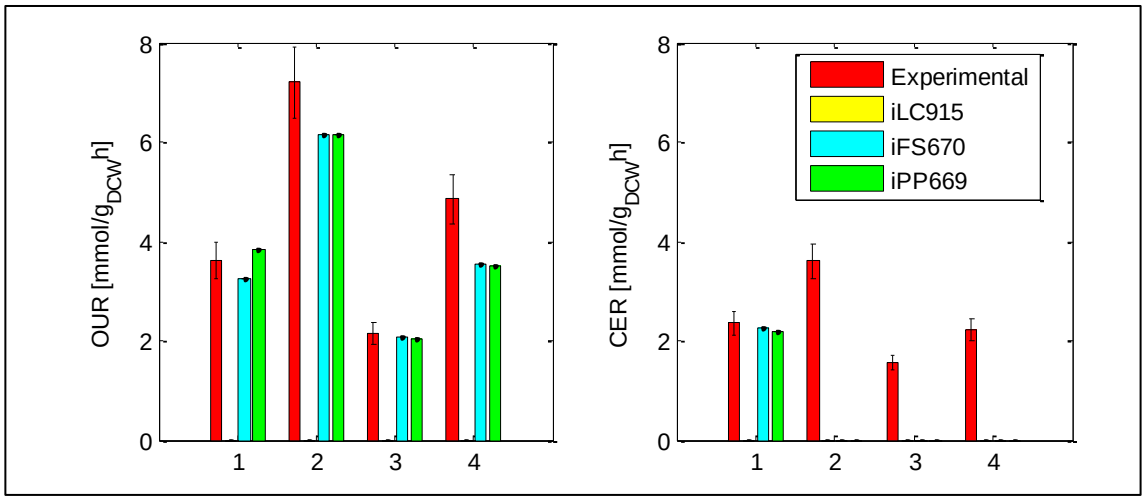


Figure 20 - Oxygen Uptake Rate (OUR) and Carbon dioxyde Evolution Rate (CER) in glycerol and methanol limited chemostats. 1: Fast growth glycerol limited chemostat, 2: Fast Growth glycerol and methanol limited chemostat, 3: Slow growth glycerol-limited chemostat and 4: Slow growth glycerol and methanol limited chemostat.

II. **Supplementary Material 2:** Feeding policies used in the optimization of bioprocesses example

We tested 13 different feeding strategies, which yielded a constant or decreasing growth rate (Figure 21). The details of the strategies are presented in Table 21.

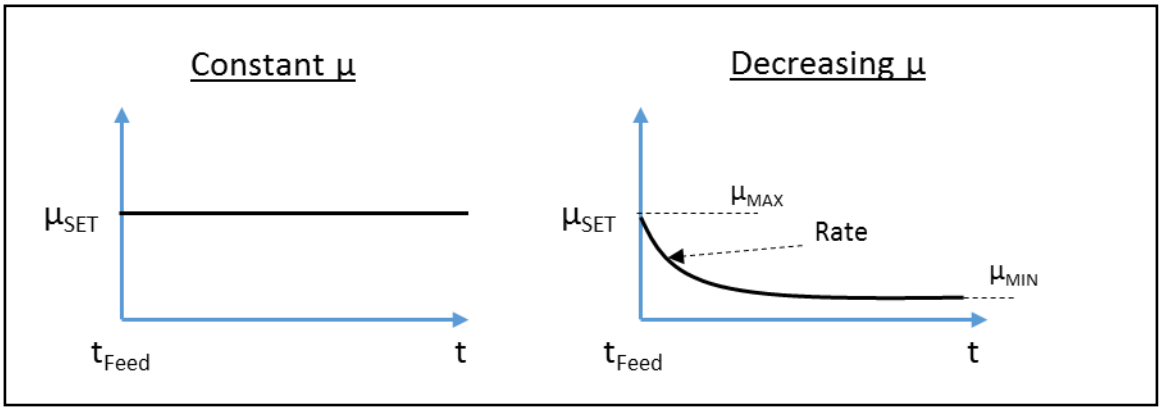


Figure 21 – Constant (left) versus decreasing (right) growth rates during fed-batch culture. Here, t_{FEED} corresponds to the time when the feed of the culture starts after batch cultivation and μ_{MAX}, Rate and μ_{MIN} refer to the parameters used to describe the decreasing growth rate profile of the culture. We evaluated two values for each one of these parameters, which yielded eight dynamic feeding strategies (6-13 in Table 21).

Table 21 - Feeding strategies evaluated and productivity indicators. The first five strategies attempt to make the culture grow at a constant growth rate while the rest produce a decreasing growth rate profile.

Strategy	$\mu_{SET,MAX}$ [h^{-1}]	Rate	$\mu_{SET,MIN}$ [h^{-1}]	q_P [$mg/g_{DCW}h$]	X_{FINAL} [g/L]	P_{FINAL} [mg/L]	Limitation
1	0,14	-	-	2,85	164,8	138	Oxygen
2	0,12	-	-	2,59	187,8	135	Oxygen
3	0,1	-	-	2,32	195,3	130	Volume
4	0,08	-	-	2,29	191,3	138	Volume
5	0,06	-	-	2,28	184,7	154	Volume
6	0,14	0,07	0,08	2,13	193,1	121,6	Volume
7	0,14	0,07	0,04	1,33	176,6	92,3	Volume
8	0,14	0,01	0,08	2,83	197,5	150,0	Volume
9	0,14	0,01	0,04	2,34	195,1	128,0	Volume
10	0,1	0,07	0,08	1,88	191,0	111,6	Volume
11	0,1	0,07	0,04	0,89	172,8	66,9	Volume
12	0,1	0,01	0,08	1,41	193,7	81,3	Volume
13	0,1	0,01	0,04	2,30	188,8	140,5	Volume

III. **Supplementary Material 3:** Batch Model Initial Calibration and Parameter Values

The model was used to calibrate data from eight aerobic, glucose-limited batch cultivations. The parameter values achieved in the calibrations is presented in Table 22, while the time of calibration and objective function value of the calibration are presented in Table 23.

Table 22 - Parameter values achieved in the calibration of data from eight batch cultivations using the initial batch model structure

	Dataset							
	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**
V_{MAX}	7.75	3.34	7.95	2.74	3.07	6.67	1.51	1.27
K_S	9.80E-04	9.60E-05	7e -4	3.00E-04	1.00E-05	1.5 e-4	1.00E-05	1.03E-05
$v_{EtOH,B}$	1.98	1.78	1.76	1.48	0.89	2.97	0.03	0.02
$v_{Pyr,B}$	0.21	0.18	0.25	0.19	0.12	0.2	0.01	0.003
$v_{Arab,B}$	0.54	0.32	0.5	0.51	0.4	0.48	0.16	0.09
$v_{Cit,B}$	0.06	0.04	0.08	1.1	0	0.05	0.03	0.05
α_B	4.05E-04	2.90E-04	4.20E-04	2.30E-04	1.45E-06	3.0E-04	7.19E-06	7.17E-05
m_{atp}	0.52	0.68	0.001	4.09	9.99	4.61	3.29	1.32

Table 23 - General features of initial model calibration.

Dataset	**N**	**Min. squares difference**	**Time of calibration [h]**
1	8	0.26	1.92
2	8	1.55	5.29
3	8	0.47	3.46
4	8	0.88	4.40
5	6	1.69	4.11
6	9	7.27	4.41
7	12	1.17	4.32
8	13	3.30	4.30

We also provide two examples of how the model fitted two of these cultivations.

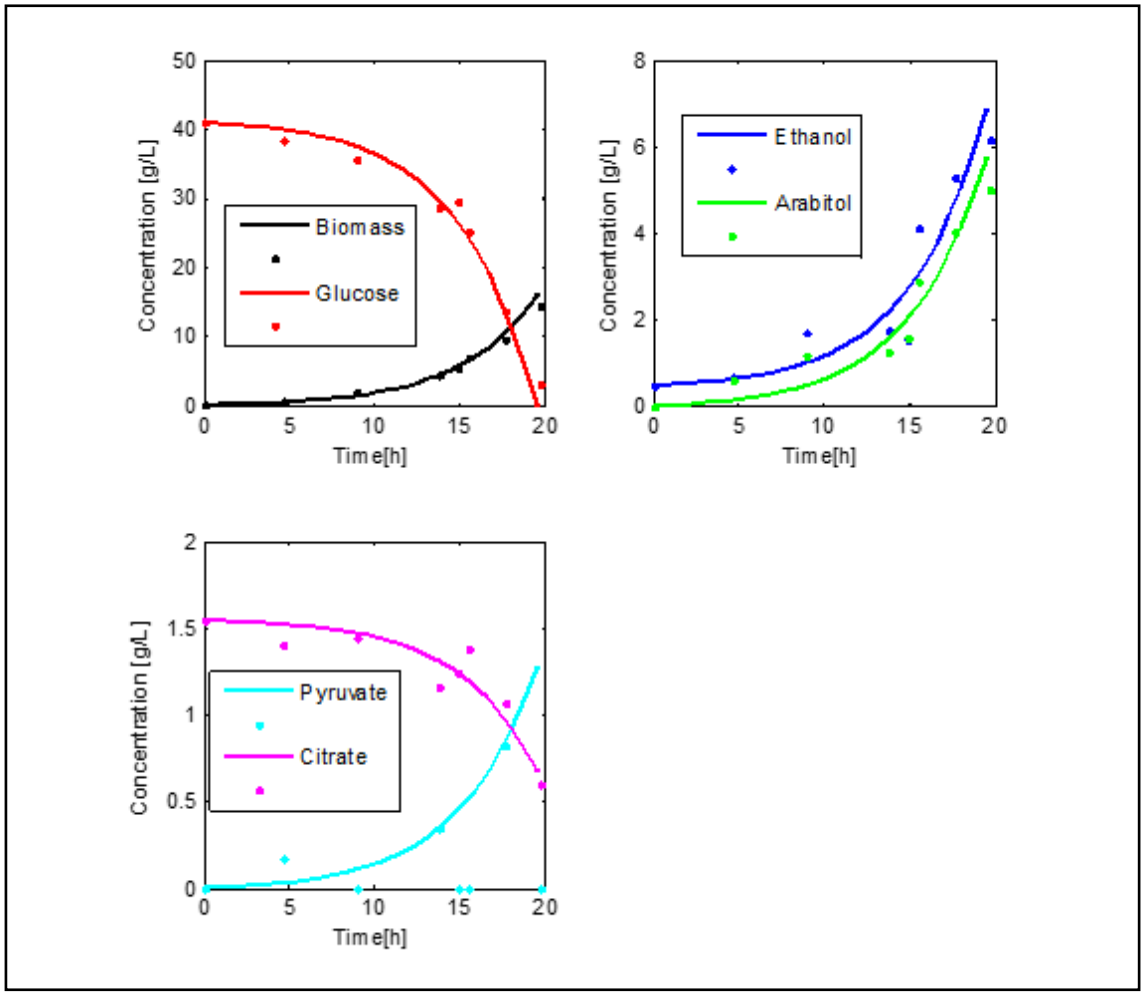


Figure 22 - Batch model calibration of GS115 culture 1

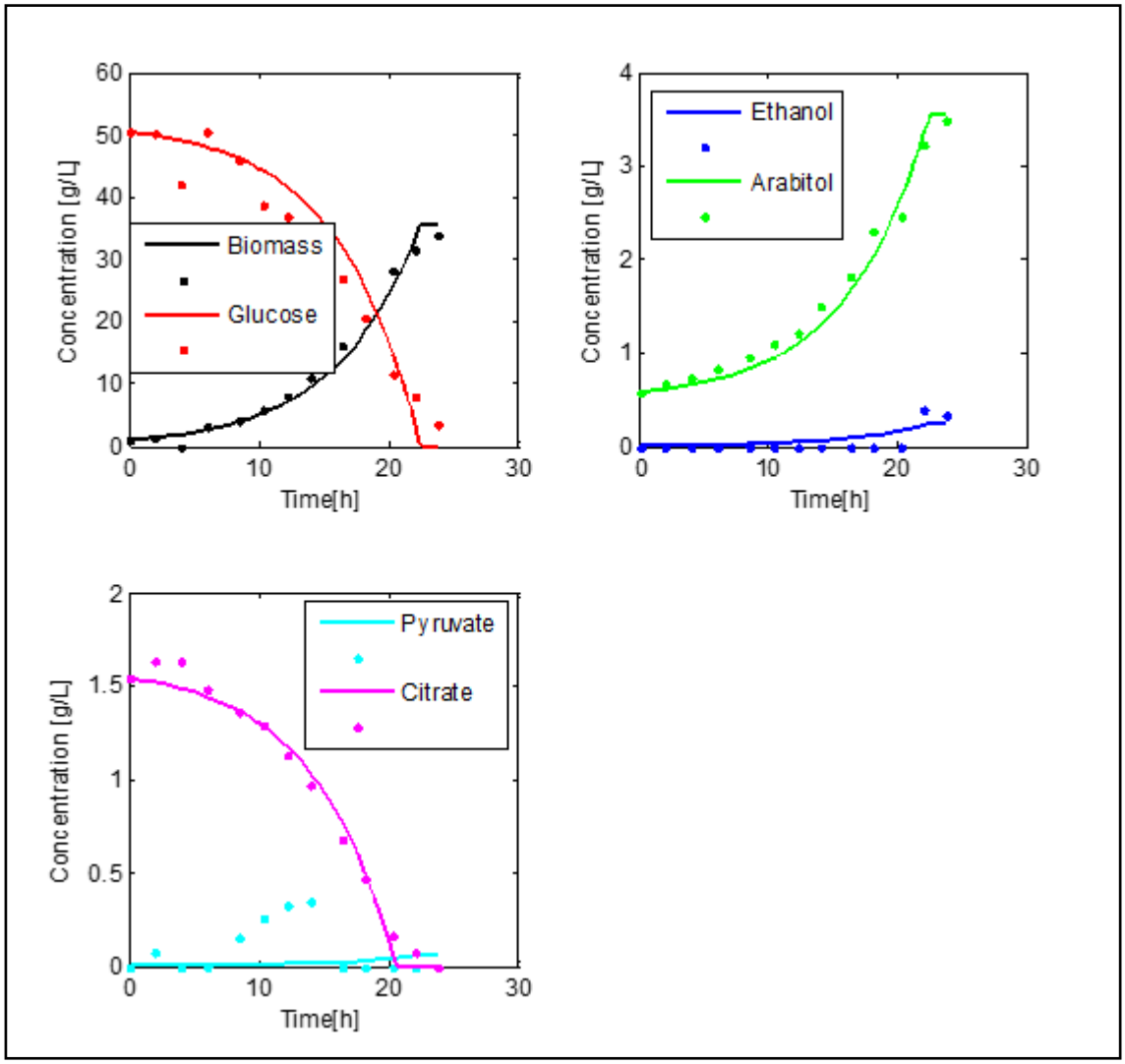


Figure 23 - Batch model calibration of GS115 culture 8

IV. **Supplementary Material 4:** Parametric issues found in the initial batch model calibrations

The recurrence of identifiability, significance and sensitivity issues found amongst the initial batch model calibration is presented in Table 24 and Table 25.

Table 24 - Percentage of calibrations (8 in total) where pairs of parameters show identifiability issues (correlation ≥ 0.95). Parameters with recurrent identifiability issues are highlighted.

	V_{MAX}	K_S	$v_{EtOH,B}$	$v_{Pyr,B}$	$v_{Arab,B}$	$v_{Cit,B}$	α_B	m_{atp}
V_{MAX}	-							
K_S	13%	-						
$v_{EtOH,B}$	50%	0%	-					
$v_{Pyr,B}$	25%	0%	38%	-				
$v_{Arab,B}$	25%	0%	25%	13%	-			
$v_{Cit,B}$	13%	0%	0%	13%	0%	-		
α_B	25%	25%	0%	0%	0%	13%	-	
m_{atp}	63%	25%	38%	25%	38%	13%	50%	-

Table 25 – Percentage (o Frequency) of calibrations (8 in total) where a parameter presented sensitivity or significance issues. Parameters with recurrent problems are highlighted.

	V_{MAX}	K_S	$v_{EtOH,B}$	$v_{Pyr,B}$	$v_{Arab,B}$	$v_{Cit,B}$	α_B	m_{atp}
Sensitivity	0%	38%	0%	0%	0%	13%	25%	13%
Significance	25%	50%	0%	13%	0%	25%	25%	25%

V. **Supplementary Material 5:** Batch model Cross Calibration Summary

Table 26 - Cross Calibration summary.

Structure	N° parameters	F_{DIFF}	% of significance Issues	% of Sensitivity Issues	% of Identifiability Issues
Original	**8**	**0**	**23.6**	**16.7**	**17.4**
1	6	-0.10	22.9	18.8	15.0
2	3	2.77	29.1	8	25
3	5	0.18	90.0	23	68
4	2	0.56	62.5	13	0
5	3	4.10	29.2	0	54
6	5	0.18	0	2.5	61.3
7	5	2.93	18.8	20.8	60.0
8	5	2.82	22.5	25.0	33.8

VI. **Supplementary Material 6:** Correlation and Sensitivity Matrixes of the calibration of the batch validation dataset

Table 27 and Table 28 show that the calibration of the validation datasets yielded no identifiability sensitivity (correlation > 0.95 between parameters) nor sensitivity (parameters without impact on the state variables) issues.

Table 27 – Correlation Matrix of the robust parameter set used to calibrate the batch validation dataset. Each cell contains the correlation between the two corresponding parameters.

	$v_{EtOH,B}$	$v_{Pyr,B}$	$v_{Arab,B}$	$v_{Cit,B}$	α_B
$v_{EtOH,B}$	1	-0,45	0,43	0,78	-0,86
$v_{Pyr,B}$	-0,45	1	-0,17	-0,56	0,50
$v_{Arab,B}$	0,43	-0,17	1	0,38	-0,43
$v_{Cit,B}$	0,78	-0,56	0,38	1	-0,88
α_B	-0,86	0,50	-0,43	-0,88	1

Table 28 - Sensitivity Matrix of the robust Parameter set used to calibrate the batch validation dataset. Each cell contains the average sensitivity of a particular parameter over the state variables.

	Volume	**Biomass**	**Glucose**	**Ethanol**	**Pyruvate**	**Arabitol**	**Citrate**
$v_{EtOH,B}$	0	0,15	0,12	0,91	0,14	0,14	0,04
$v_{Pyr,B}$	0	0,02	0,00	0,02	0,95	0,02	0,00
$v_{Arab,B}$	0	0,03	0,01	0,02	0,02	0,97	0,00
$v_{Cit,B}$	0	0,03	0,00	0,02	0,03	0,03	0,45
α_B	0	2,62	1,09	1,80	1,93	2,00	0,73

VII. **Supplementary Material 7:** Fed-batch model initial Calibration and Parameter Values

Table 29 indicates the parameter values achieved for the three fed-batch cultivations used in the initial calibration of the fed-batch model. Figures 24, 25 and 26, show the model fits to the experimental data.

Table 29 - Parameter Values of the Initial calibrations performed with the complete fed-batch model (14 parameters)

Parameter	Dataset 1	Dataset 2	Dataset 3	Mean	Units
$v_{S,max}$	2,74	3,29	2,59	2,94	$mmol/g_{DCW}h$
K_S	0,05	0,03	0,07	0,05	g/L
$v_{EtOH,B}$	1,95	2,18	0,98	1,58	$mmol/g_{DCW}h$
$v_{Pyr,B}$	0,18	0,18	0,13	0,15	$mmol/g_{DCW}h$
$v_{Arab,B}$	0,5	0,24	0,11	0,18	$mmol/g_{DCW}h$
$v_{Cit,B}$	0,12	0,11	0,22	0,16	$mmol/g_{DCW}h$
$v_{EtOH,FB}$	1,13	1,20	1,22	1,21	$mmol/g_{DCW}h$
$v_{Pyr,FB}$	0,10	0,02	0,26	0,14	$mmol/g_{DCW}h$
$v_{Arab,FB}$	0,07	0,13	0,17	0,15	$mmol/g_{DCW}h$
$v_{Cit,FB}$	0,005	0,00	0,01	0,008	$mmol/g_{DCW}h$
α_B	3,03E-04	4,26E-05	1,49E-04	9,6E-05	$[-]$
α_{FB}	1,28E-04	2,22E-14	1,44E-04	7,2E-05	$[-]$
m_{ATP}	4,38	9,00	8,13	8,6	$mmol/g_{DCW}h$
T_{Fed}	23	22,02	22,94	22,5	h

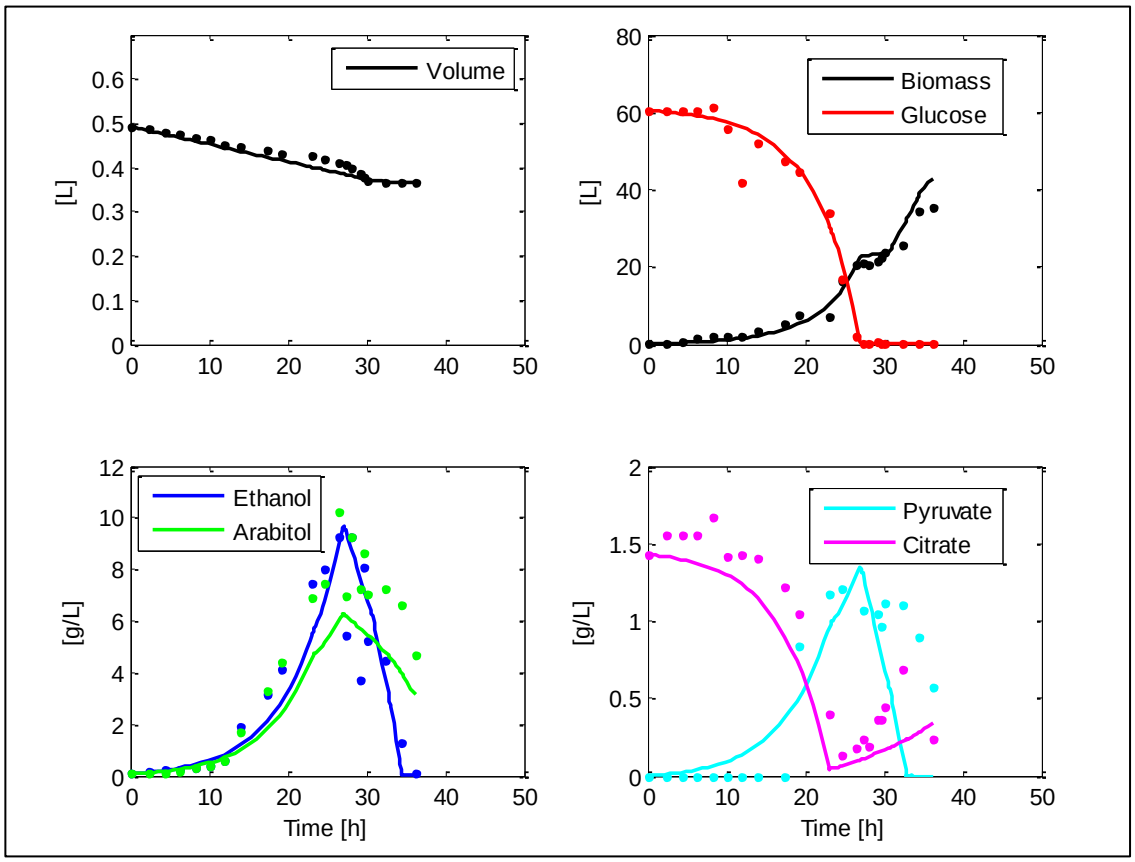


Figure 24 - Calibration of fed-batch dataset 1 using the original model structure

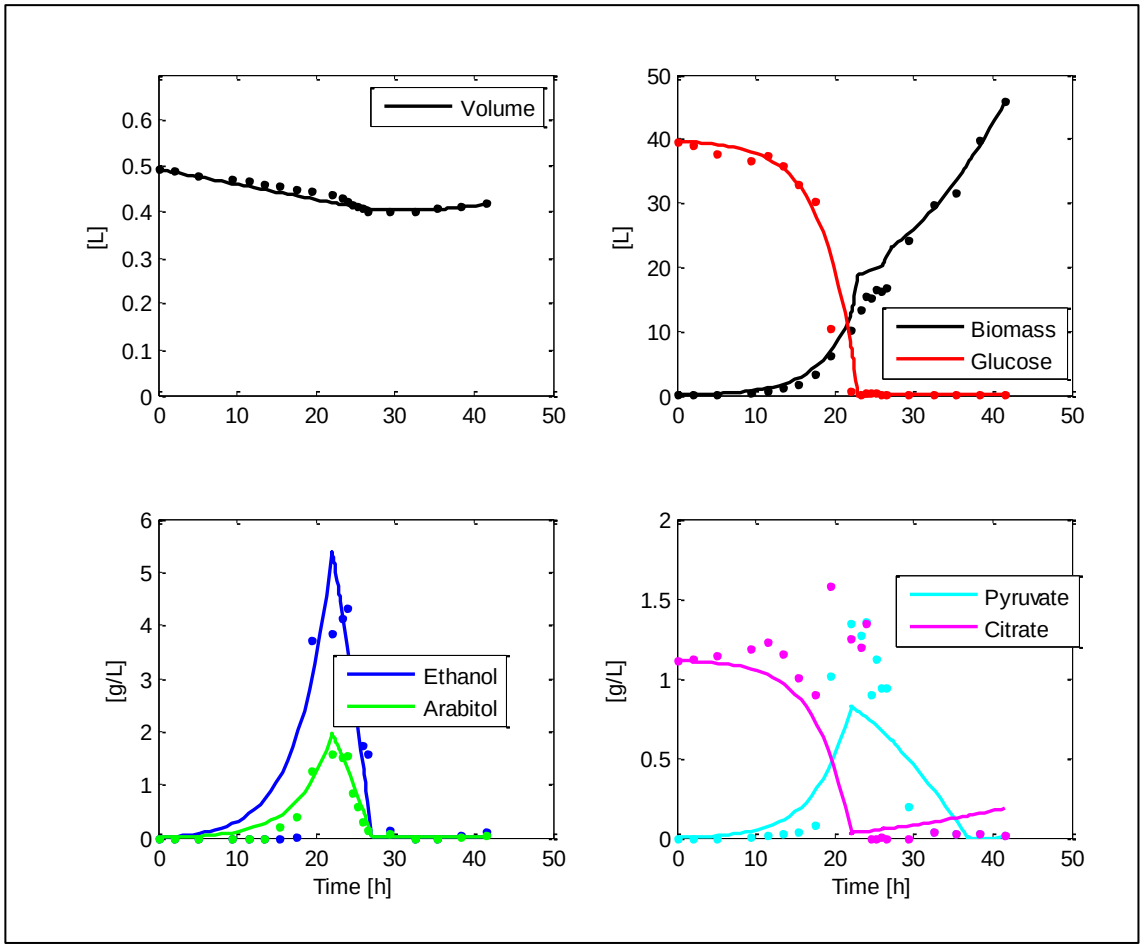


Figure 25 - Calibration of fed-batch dataset 2 using the original model structure

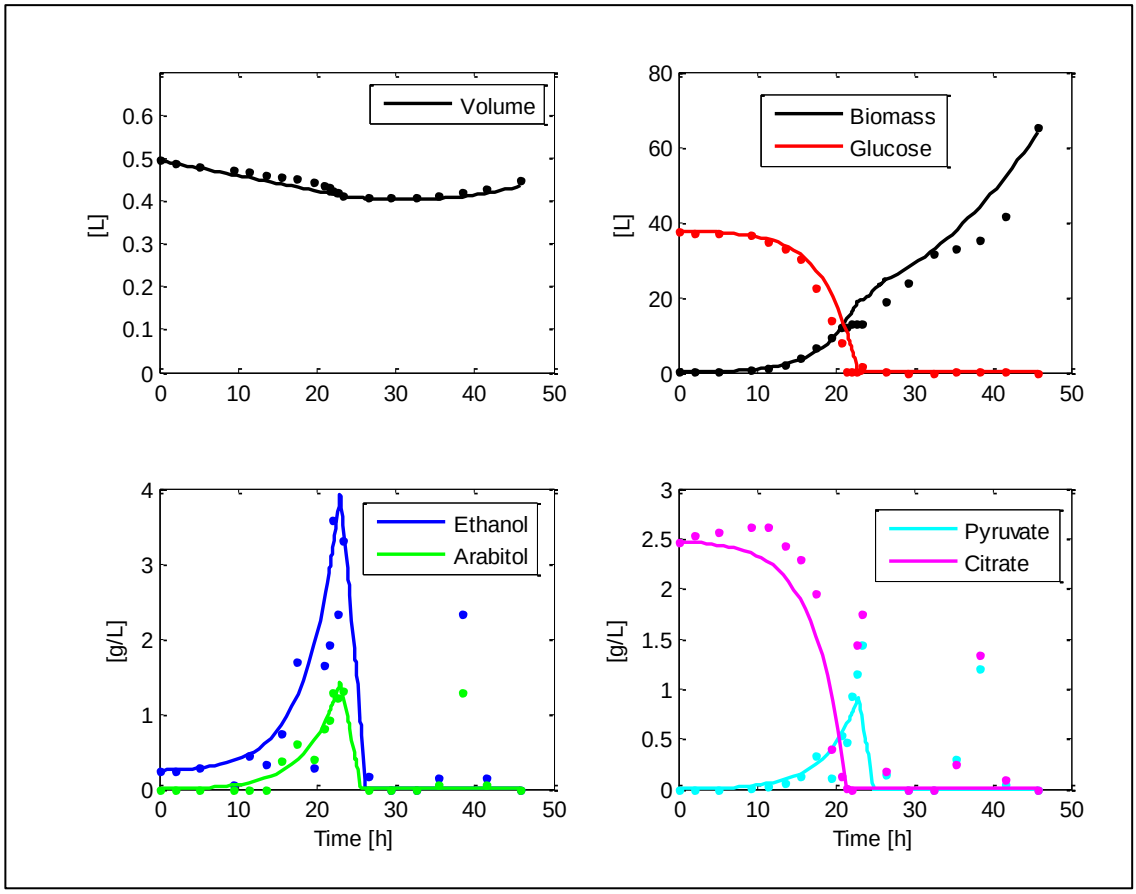


Figure 26 - Calibration of fed-batch dataset 3 using the original model structure

VIII. **Supplementary Material 8:** Correlation and Sensitivity Matrixes of the fed-batch validation dataset calibrated with the candidate robust modeling structure 3.

Table 30 - Correlation Matrix of the calibration of the fed-batch validation dataset

	v_{MAX}	K_S	$v_{Pyr,B}$	$v_{Arab,B}$	$v_{Cit,B}$	$v_{Pyr,FB}$	α_B	α_{FB}	m_{ATP}	T_{Cons}
v_{MAX}	1	0,15	0,17	-0,47	0,45	-0,11	-0,92	0,14	-0,85	0,04
K_S	0,15	1	0,94	-0,85	0,62	-0,49	-0,20	0,68	0,24	-0,32
$v_{Pyr,B}$	0,17	0,94	1	-0,86	0,67	-0,76	-0,20	0,89	0,25	-0,42
$v_{Arab,B}$	-0,47	-0,85	-0,86	1	-0,66	0,56	0,44	-0,70	0,07	0,55
$v_{Cit,B}$	0,45	0,62	0,67	-0,66	1	-0,53	-0,51	0,60	-0,20	-0,10
$v_{Pyr,FB}$	-0,11	-0,49	-0,76	0,56	-0,53	1	0,15	-0,96	-0,19	0,42
α_B	-0,92	-0,20	-0,20	0,44	-0,51	0,15	1	-0,18	0,88	-0,08
α_{FB}	0,14	0,68	0,89	-0,70	0,60	-0,96	-0,18	1	0,21	-0,42
m_{ATP}	-0,85	0,24	0,25	0,07	-0,20	-0,19	0,88	0,21	1	-0,25
T_{Cons}	0,04	-0,32	-0,42	0,55	-0,10	0,42	-0,08	-0,42	-0,25	1

Table 31 - Sensitivity Matrix of the calibration of the fed-batch validation dataset

	Volume	**Biomass**	**Glucose**	**Ethanol**	**Pyruvate**	**Arabitol**	**Citrate**
v_{MAX}	0,00	0,65	0,93	4,32	3,67	1,57	0,61
K_S	0,00	0,22	0,00	0,01	1,24	0,02	0,00
$v_{Pyr,B}$	0,00	0,15	0,00	0,00	1,02	0,00	0,00
$v_{Arab,B}$	0,00	0,10	0,09	0,69	2,74	0,91	0,22
$v_{Cit,B}$	0,00	0,03	0,00	0,09	0,96	0,04	0,62
$v_{Pyr,FB}$	0,00	0,13	0,00	0,00	0,24	0,03	0,00
α_B	0,00	1,03	0,88	5,82	6,27	2,91	1,24
α_{FB}	0,00	0,03	0,00	0,01	0,00	0,02	0,00
m_{ATP}	0,00	0,79	0,65	3,24	3,66	2,13	0,95
T_{Cons}	0,00	0,26	0,00	0,22	3,66	3,06	0,88

IX. **Supplementary Material 9:** Glossary for Figure 12, analysis of the metabolic flux distribution throughout a fed-batch Cultivation.

The abbreviations used correspond to the following metabolites:

- G6P = Glucose 6 Phosphate
- Ru5P = Ribulose 5 Phosphate
- ABT = Arabitol
- PPP = Non-oxidative phase of the Pentose Phosphate Pathway
- F6P = Fructose 6 Phosphate
- G3P = Glyceraldehyde 3 Phosphate
- DHAP = Dihydroxyacetone Phosphate
- Pyr = Pyruvate
- OAA = Oxaloacetate
- Acald = Acetaldehyde
- EtOH = Ethanol
- AcCoA = Acetyl Coenzyme A
- Cit = Citrate
- Icit = Isocitrate
- αkg = Alpha-keto glutarate
- Mal = Malate
- L- glut = Glutamate

X. **Supplementary Material 10:** Knockout candidates derived by MOMA

Table 32 - Knockout candidates for HSA overproduction

	Deleted Gene	Final Biomass [g/L]	Final HSA [g/L]	Reaction(s) Name(s)
1	PAS_chr2-2_0094	11,81	0,91	Chitin synthase
2	PAS_chr1-4_0194	11,41	0,84	Putrescine and Spermidine transport
3	PAS_chr4_0836	11,41	0,84	Putrescine and Spermidine transport
4	RPPA11109	10,74	0,63	ribulose 5-phosphate 3-epimerase
5	RPPA11110	10,74	0,63	ribulose 5-phosphate 3-epimerase
6	PAS_chr2-2_0330	10,20	0,62	Phosphoryl ceramide syntase
7	PAS_chr2-2_0044	8,59	0,30	CDP-Diacylglycerol synthetase, yeast-specific
8	PAS_chr4_0210	8,84	0,29	ADP/ATP transporter, mitochondrial
9	PAS_chr4_0212	8,84	0,29	ribose-5-phosphate isomerase
10	PAS_chr3_0604	14,05	0,23	Deoxyribokinase and ribokinase
11	PAS_chr4_0408	15,33	0,22	phosphoethanolamine cytidyltransferase
12	PAS_chr1-1_0418	16,02	0,21	Acetate transporter
13	PAS_chr1-3_0220	15,88	0,19	Methylenetetrahydrofolate dehydrogenase NAD
14	PAS_chr1-4_0487	14,85	0,17	Succinate Dehydrogenase
15	PAS_chr2-2_0278	14,85	0,17	Peptide alpha-N-acetyltransferase
16	PAS_chr3_1110	14,85	0,17	Tyrosyl-tRNA synthetase, mitochondrial
17	PAS_chr4_0733	14,85	0,17	Succinate Dehydrogenase
18	PAS_chr3_0646	14,04	0,14	Phospholipase D, yeast-specific
19	PAS_chr3_0471	15,44	0,12	aspartate-semialdehyde dehydrogenase, irreversible
20	PAS_chr2-1_0657	13,40	0,12	phosphoglycerate dehydrogenase
21	PAS_chr4_0284	13,40	0,12	ribonucleoside-diphosphate reductase
22	PAS_chr4_0877	16,73	0,05	malate, succinate and fumarate transport, mitochondrial
23	PAS_chr3_0176	16,53	0,05	N-acteylglutamate synthase and ornithine transacetylase , mitochondrial
24	PAS_chr1-1_0050	9,78	0,05	Pyruvate dehydrogenase
25	PAS_chr1-4_0254	9,78	0,05	Ppyruvate dehydrogenase
26	PAS_chr1-4_0593	9,78	0,05	Pyruvate dehydrogenase, tetrahydrofolate aminomethyltransferase
27	PAS_chr2-2_0288	9,78	0,05	Arginase
28	PAS_chr3_0649	16,26	0,03	Thiamine transport in via proton symport
29	PAS_chr2-2_0127	17,10	0,03	Cytochrome c peroxidase, mitochondrial
30	PAS_chr1-4_0659	17,10	0,03	Hydrogen peroxide reductase thioredoxin, peroxisomal
31	PAS_chr2-1_0547	16,70	0,03	3',5'-bisphosphate nucleotidase

32	PAS_chr3_0462	17,14	0,03	Alanyl-tRNA synthetase

Table 33 - Reactions and pathways associated to the deletion candidates

	Deleted Gene	Reactions	Pathway
1	PAS_chr2-2_0094	udpacgam[c] => h[c] + udp[c] + chitin[c]	Glutamate metabolism
2	PAS_chr1-4_0194	h[c] + ptrc[e] => h[e] + ptrc[c] h[c] + spmd[e] => h[e] + spmd[c] h[c] + sprm[e] => h[e] + sprm[c]	Transport, Extracellular
3	PAS_chr4_0836	h[c] + ptrc[e] => h[e] + ptrc[c] h[c] + spmd[e] => h[e] + spmd[c] h[c] + sprm[e] => h[e] + sprm[c]	Transport, Extracellular
4	RPPA11109	ru5p-D[c] <=> xu5p-D[c]	Pentose Phosphate Pathway
5	RPPA11110	ru5p-D[c] <=> xu5p-D[c]	Pentose Phosphate Pathway
6	PAS_chr2-2_0330	ptd1ino_PP[c] + cer1_24[c] => 12dgr_PP[c] + ipc124_PP[c]	Sphingolipid Metabolism
7	PAS_chr2-2_0044	h[c] + pa_PP[c] + ctp[c] <=> ppi[c] + cdpdag_PP[c] h[m] + ctp[m] + pa_PP[m] <=> ppi[m] + cdpdag_PP[m]	Phospholipid Biosynthesis
8	PAS_chr4_0210	h[c] + adp[c] + atp[m] => h[m] + atp[c] + adp[m]	Transport, Mitochondrial
9	PAS_chr4_0212	r5p[c] <=> ru5p-D[c]	Pentose Phosphate Pathway
10	PAS_chr3_0604	atp[c] + rib-D[c] => h[c] + adp[c] + r5p[c]	Pentose Phosphate Pathway
11	PAS_chr4_0408	h[c] + ctp[c] + ethamp[c] => ppi[c] + cdpea[c]	Phospholipid Biosynthesis
12	PAS_chr1-1_0418	ac[e] <=> ac[c]	Transport, Extracellular
13	PAS_chr1-3_0220	nad[c] + mlthf[c] => nadh[c] + methf[c]	Folate Metabolism
14	PAS_chr1-4_0487	fad[m] + succ[m] <=> fadh2[m] + fum[m] q6[m] + succ[m] <=> q6h2[m] + fum[m] q6[m] + fadh2[m] <=> q6h2[m] + fad[m]	Citric Acid Cycle/Oxydative Phosphorilation
15	PAS_chr2-2_0278	accoa[c] + pepd[c] => h[c] + coa[c] + apep[c]	Other Amino Acid Metabolism
16	PAS_chr3_1110	atp[m] + tyr-L[m] + trnatyr[m] => amp[m] + ppi[m] + tyrtrna[m]	tRNA charging
17	PAS_chr4_0733	fad[m] + succ[m] <=> fadh2[m] + fum[m] q6[m] + succ[m] <=> q6h2[m] + fum[m] q6[m] + fadh2[m] <=> q6h2[m] + fad[m]	Citric Acid Cycle/Oxydative Phosphorilation
18	PAS_chr3_0646	h2o[c] + pc_PP[c] => h[c] + pa_PP[c] + chol[c]	Phospholipid Metabolism
19	PAS_chr3_0471	h[c] + nadph[c] + 4pasp[c] => pi[c] + nadp[c] + aspsa[c]	Alanine and Aspartate Metabolism
20	PAS_chr2-1_0657	nad[c] + 3pg[c] => h[c] + nadh[c] + 3php[c]	Glycine and Serine Metabolism
21	PAS_chr4_0284	19 Reactions	Nucleotide Salvage Pathway

22	PAS_chr4_0877	pi[m] + mal-L[c] <=> pi[c] + mal-L[m]	Transport, Mitochondrial
23	PAS_chr3_0176	accoa[m] + glu-L[m] => h[m] + coa[m] + acglu[m] glu-L[m] + acorn[m] => acglu[m] + orn[m]	Arginine and Proline Metabolism
24	PAS_chr1-1_0050	nad[m] + coa[m] + pyr[m] => nadh[m] + co2[m] + accoa[m]	Glycolysis/Gluconeogenesis
25	PAS_chr1-4_0254	nad[m] + coa[m] + pyr[m] => nadh[m] + co2[m] + accoa[m]	Glycolysis/Gluconeogenesis
26	PAS_chr1-4_0593	udpacgam[c] => h[c] + udp[c] + chitin[c]	Glycolysis/Gluconeogenesis
27	PAS_chr2-2_0288	h[c] + ptrc[e] => h[e] + ptrc[c] h[c] + spmd[e] => h[e] + spmd[c] h[c] + sprm[e] => h[e] + sprm[c]	Arginine and Proline Metabolism
28	PAS_chr3_0649	h[c] + ptrc[e] => h[e] + ptrc[c] h[c] + spmd[e] => h[e] + spmd[c] h[c] + sprm[e] => h[e] + sprm[c]	Transport, Extracellular
29	PAS_chr2-2_0127	ru5p-D[c] <=> xu5p-D[c]	Oxidative Phosphorylation
30	PAS_chr1-4_0659	ru5p-D[c] <=> xu5p-D[c]	Other
31	PAS_chr2-1_0547	ptd1ino_PP[c] + cer1_24[c] => 12dgr_PP[c] + ipc124_PP[c]	Cysteine Metabolism
32	PAS_chr3_0462	h[c] + pa_PP[c] + ctp[c] <=> ppi[c] + cdpdag_PP[c] h[m] + ctp[m] + pa_PP[m] <=> ppi[m] + cdpdag_PP[m]	tRNA charging

XI. **Supplementary material 11:** Similarity to experimental data of the iMT1026 model and a batch calibration using it

The iMT1026 model has better predicting capability of experimental consumption and production rates compared to the iFS670 model (Table 34).

Table 34 - Average error associated to experimental rate predictions of the iFS670 model versus the iMT1026 model

Carbon Source	**N**	**iFS670**	**iMT1026**
Glycerol/MetOH	12	37%	24%
Glucose	30	36%	31%
Average Error F (%) =	42	36%	29%

As the new model improved the predictions, we used it to calibrate a batch cultivation assuming biomass maximization. The metabolic block did not solve a QP problem and the model contained six parameters (Table 35 and Figure 27).

Table 35 - Parameter values achieved in the calibration of a batch cultivation using the iMT1026 genome-scale metabolic model

Parameter	**Value**	**Units**
V_{MAX}	3.56	mmol/g_{DCW}h
K_S	0.008	g/L
$v_{EtOH,B}$	2.00	mmol/g_{DCW}h
$v_{Pyr,B}$	0.22	mmol/g_{DCW}h
$v_{Arab,B}$	0.57	mmol/g_{DCW}h
$v_{Cit,B}$	0.07	mmol/g_{DCW}h

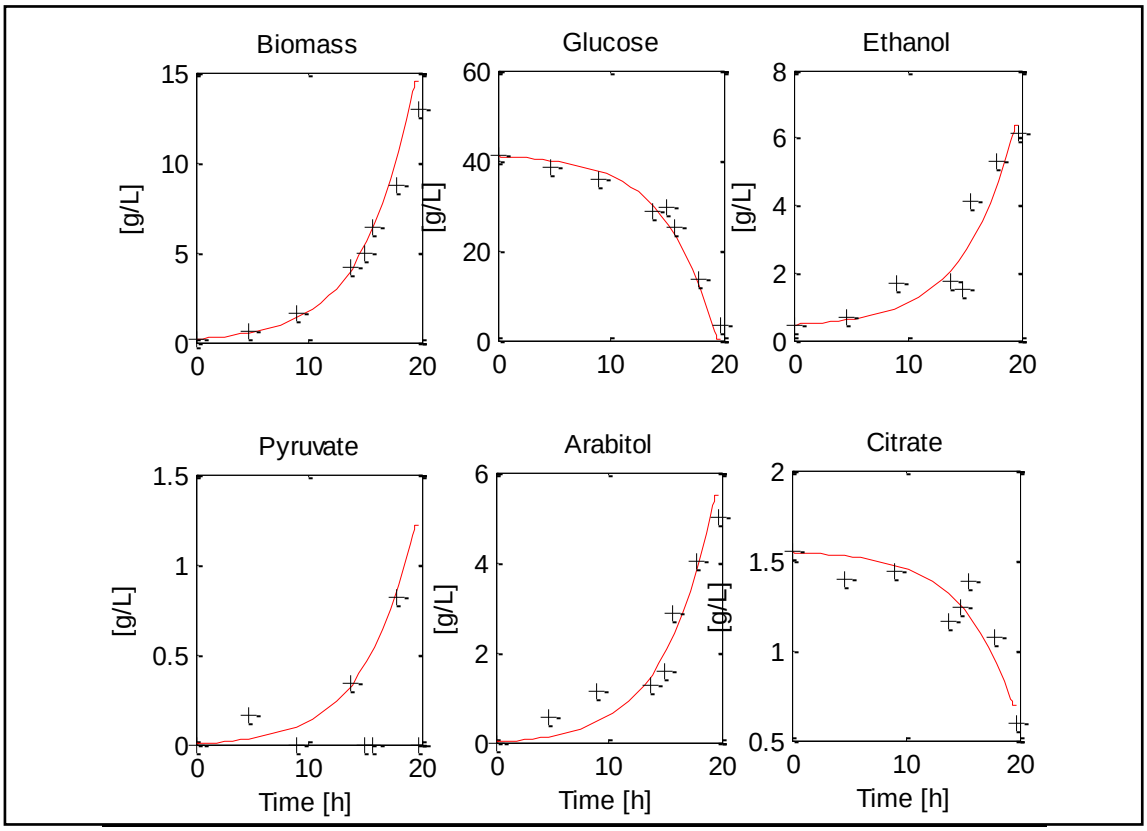


Figure 27 - Calibration of a batch cultivation using the iMT1026 genome-scale model of *Pichia pastoris*