González Martínez, JM.; Abel Folch-Fortuny; Llaneras Estrada, F.; Tortajada Serra, M.; Picó Marco, JA.; Ferrer, A. (2014). Metabolic flux understanding of Pichia pastoris grown on heterogenous culture media. Chemometrics and Intelligent Laboratory Systems. 134:89-99. doi:10.1016/j.chemolab.2014.02.003.

Metabolic Flux Understanding of *Pichia pastoris* Grown on Heterogenous Culture Media

J. M. González-Martínez[a,b,*], A. Folch-Fortuny[b], F. Llaneras[c], M. Tortajada[d], J. Picó[e], A. Ferrer[b]

[a] *Shell Global Solutions International B.V., Shell Technology Centre Amsterdam, PO Box 38000, 1030 BN Amsterdam, The Netherlands*
[b] *Departamento de Estadística e Investigación Operativa Aplicadas y Calidad, Universitat Politècnica de València, Camino de Vera s/n, Edificio 7A, 46022 Valencia, Spain*
[c] *Department of Electrical, Electronic and Automatic Engineering, Universitat de Girona, Campus de Montilivi, 17071 Girona, Spain*
[d] *Biopolis, S.L., Parc Científic Universitat de València, c/ Catedrático Agustín Escardino Benlloch 9, 46980 Paterna, Spain*
[e] *Institut Universitari d´ Automàtica i Informàtica Industrial, Universitat Politècnica de València, Camino de Vera s/n, Edificio 5C, 46022 Valencia, Spain*

Abstract

Within the emergent field of Systems Biology, mathematical models obtained from physical-chemical laws (the so-called first principles-based models) of microbial systems are employed to discern the principles that govern cellular behaviour and achieve a predictive understanding of cellular functions. The reliance on this biochemical knowledge has the drawback that some of the assumptions (specific kinetics of the reaction system, unknown dynamics and values of the model parameters) may not be valid for all the metabolic possible states of the network. In this uncertainty context, the combined use of fundamental knowledge and data measured in the fermentation that describe the behavior of the microorganism in the manufacturing process is paramount to overcome this problem. In this paper, a grey modelling approach is presented combining data-driven and first principles information at different scales, developed for *Pichia pastoris* cultures grown on different carbon sources. This approach will allow us to relate patterns of recombinant protein production to intracellular metabolic states and correlate intra and extracellular reactions in order to understand how the internal state of the cells determines the observed behaviour in *P. pastoris* cultivations.

Keywords: *Metabolic network*, *Possibilistic consistency anlaysis*, *Monte Carlo sampling*, *Principal Component Analysis*, *Missing-data methods for Exploratory Data Analysis*

1. Introduction

Currently, biotechnological industries are devoted to the production of economically important enzymes and proteins, generally using genetically modified microorganisms. The main goal of these industries is to maximize protein yields and productivity. The production of these high-added value products is governed by highly correlated factors that require a multidisciplinary approach to process optimization (biochemistry, molecular biology, process engineering, biotechnology, *etc*).

The development of accurate monitoring schemes to control the manufacturing process becomes a challenging task due to the scarcity of measurements and the high complexity of the biochemical synthesis process. Only few process variables can be measured in industrial microbial fermentations, such as pH, temperature, and oxygen consumption. Others, such as substrate consumption, can be inferred depending on the operational strategy. In this kind of processes, measurements corresponding to biological process variables (such as intracellular specific reaction rates) are promising to achieve a more accurate process

*Corresponding author.
Email address: jogonmar@gmail.com (J. M. González-Martínez)

control. In order to develop novel monitoring schemes, the study of different cellular behaviours is crucial for the biotechnological production of high-added value biochemicals. For this purpose, the modelling of the available data is needed to know which key variables control the main metabolic pathways and, possibly, their regulation mechanisms.

First principles-based models of microbial systems can be developed to describe the principles that govern cellular behaviour and achieve a predictive understanding of cellular functions [1]. Typically, networks of biochemical reactions are used to approach an organism microbial metabolism and growth [2, 3]. These networks are modelled assuming that certain constrains operate at steady-state, such as environmental constraints [4], regulatory constraints [5, 6], gene expression data [7], mass balances or reactions irreversibilities [8] (the so-called *constraint-based perspective*) [9, 10, 11]. The imposed constraints define a solution space that encloses all the possible states of the network (*i.e.* flux distributions through the reactions). The development of this type of models based solely on fundamental or knowledge information has the drawback that the unknown part of the process is not represented as well as some of the underlaying assumptions (*e.g.* specific kinetics of the reaction system, unknown dynamics, values of the model parameters, objective functions) may not be valid for all the metabolic possible states of the network [12, 13]. To address this problem, hybrid models that combine knowledge-based models, which fit the theoretical behavior, and empirical models, which fit any remaining systematic variation, can be used [14].

In the context of grey modelling, there are different approaches to descompose the data into the three types of variation (known causes, unknown causes and residuals) [15], which be roughly classified into three categories. The first category are the models based on known constraints. There exist general frameworks that enable to impose very specific constraints on each type of information, *e.g.* observed experimental information [16] or transformations on the original variables [17]. These methods are based on the projection of a data matrix, followed by multivariate model decomposition. Principal Component Analysis (PCA) [18] is one of the most applied multivariate statistical projection methods to reveal the internal structure of the cell. This analysis is commonly preceded by a Monte Carlo sampling in order to produce a data set of possible states or feasible solutions from which the PCA elucidates the meaningful principal components (PCs) [19, 20, 21]. PCA has also been compared to other multivariate techniques, such as Multivariate Linear Regression [21] and Parallel Factor Analysis (PARAFAC) [22, 23] in the field of Systems Biology. Partial Least Squares regression (PLS) [24] has been applied directly [25, 26] and combined with Hierarchical Clustering (HC-PLSR) [27] to deal with situations where the input-output relations (e.g. the effect of the substrates consumption of the cell or the environmental conditions in the production of a particular protein) are highly nonlinear or non-monotone. Recently, Grey Component Analysis (GCA) has been proposed using a cost function to maximise the interpretability of the solutions by forcing the decomposition towards the direction of the prior information - a chemically or biologically meaningful solution - [28]. A second strategy is formed by methods based on introducing *a priori* knowledge by means of mathematical relations that describe the system behaviour or dynamics. The starting point is some specific structure based on first principles mathematical relations, where some functions must be estimated. Different tools can be used to calculate these functions, such as artificial Neural Networks (NNs) [29] or Kalman filters [30, 31]. Finally, a third category are the methods based on incorporating the fundamental knowledge through constraints on the modelling algorithms. For instance, some model parameters can be forced to have values within certain regions in the parameters space [32]. Projection to Latent Pathways (PLP) [33] has been recently formulated as a modification of the PLS regression algorithm by using the concept of Elementary Modes (*i.e.* thermodinamically feasible pathways through the metabolic network). This method is devoted to obtain a more biologically explanatory set of latent variables (LVs) relating the observed behaviour of the cell and its initial conditions.

The complexity of data available from microbial systems requires the design of sophisticated grey models that combine data-driven and knowledge-based information at different scales for biochemical process understanding. The main goal of this paper is to use this hybrid framework to analyse the behaviour of the methylotrophic yeast *P. pastoris* [3], as a first step to analysing which conditions and through which reactions the cell achieves an optimal state for our interests. Several scenarios corresponding to different chemostat runs are collected from the literature [34, 35, 36, 37, 38, 39, 40, 41, 42] with the aim of starting the analysis with a rich data set of different culture conditions. A recently developed adaptation [3, 43]

of the possibilistic theory [44] is applied in order to check the consistency between model and data. For the completion of the unmeasured data, a Monte Carlo sampling method is applied to produce feasible flux solutions for the microbial system under study. At this point, a PCA is performed to obtain a reduced number of PCs explaining most of the variance of the collected and sampled data. Finally, the Missing-data method for Exploratory Data Analysis (MEDA) [45] is applied to obtain a better interpretation of the PCs derived from the PCA model.

This paper is organized as follows. Section 2 presents the metabolic network reconstruction of the yeast *P. pastoris* and the different scenarios used in the study. Section 3 describes the grey modelling approach proposed in detail. This procedure is applied to the available data from *P. pastoris* in Section 4. Finally, some conclusions on the grey modelling approach presented in this paper and how it may be applied to improve the understanding of microbial cultures are drawn in Section 5.

2. Materials

Metabolic network reconstruction

The methylotrophic *P. pastoris* has become one of the most widely used yeasts for heterologous protein production since its development, in the early 1970s [46]. This system is of particular industrial interest due to its powerful and tightly regulated methanol-inducible alcohol oxidase 1 promoter (pAOX1), its capacity for foreign protein secretion, its ability to perform post-translational modifications (including glycosylation and disulfide bond formation) and its capability to grow on defined media at high cell densities [47, 48].

The constraint-based model, whose corresponding metabolic network is shown in Figure 1, has been used throughout this work. The model is a simplified representation of the whole metabolism of the yeast *P. pastoris*, meaning that only a reduced number of biochemical reactions has been included (45), from the larger amount available from genomic information (more than 1200). The reactions were selected on the basis of previous models found in literature, as lumped equivalents of more complex pathways. This model was previously validated by the authors for this organism and the experimental conditions studied [3, 49] and is the only one used in the referred experiments throughout this work. The model represents the most significant features of *P. pastoris* metabolism, including the main catabolic pathways of the yeast, such as glycolysis, the citric acid cycle, glycerol and methanol oxidation and fermentative pathways [49]. Anabolism is introduced through the pentose phosphate pathway and a general lumped biomass equation according to which growth is assumed to depend exclusively on key biochemical precursors. Branch-point metabolites, such as NADH, NADPH, AcCoA, oxalacetate and pyruvate, are considered in compartmentalized cytosolic and mitochondrial pools [34].

P. pastoris experimental data set

In this work, experimental data from several fermentation runs with different *P. pastoris* strains have been taken from the available literature, building the different scenarios considered for the subsequent statistical analysis. For the sake of visualization, the 40 scenarios under study have been grouped attending to the experimental substrates (*i.e. glucose*, *glycerol*, *methanol*, and *glycerol* and *methanol* mixtures) (see Figure 2). Scenario *A1* is taken from the strain expressing the Fab fragment of the human anti-HIV antibody 3H6 [34]. Scenarios from *B1* to *B7*, and *C1* and *C2*, are from a strain expressing a *Rhizopus oryzae* lipase (ROL) [35, 36]. Scenarios from *D1* to *D10* come from a *P. pastoris* strain expressing and secreting recombinant avidin [37]. Scenario *E1* has been obtained from a macrokinetic model for *P. pastoris* expressing recombinant human serum albumin (HSA) [38]. Scenarios from *F1* to *F7* are from a *P. pastoris* strain genetically modified to produce sea raven antifreeze protein [39]. Scenarios from *G1* to *G10* are obtained from a *P. pastoris* strain producing recombinant human chymotrypsinogen B [40]. Scenario *H1* has been obtained from the continuous fermentation of a *P. pastoris* strain for the extracellular production of a recombinant ovine interferon protein [41]. Finally, scenario *I1* comes from the expression of recombinant chitinase with a genetically modified *P. pastoris* strain [42]. The data for all these scenarios are detailed in Figure 2.

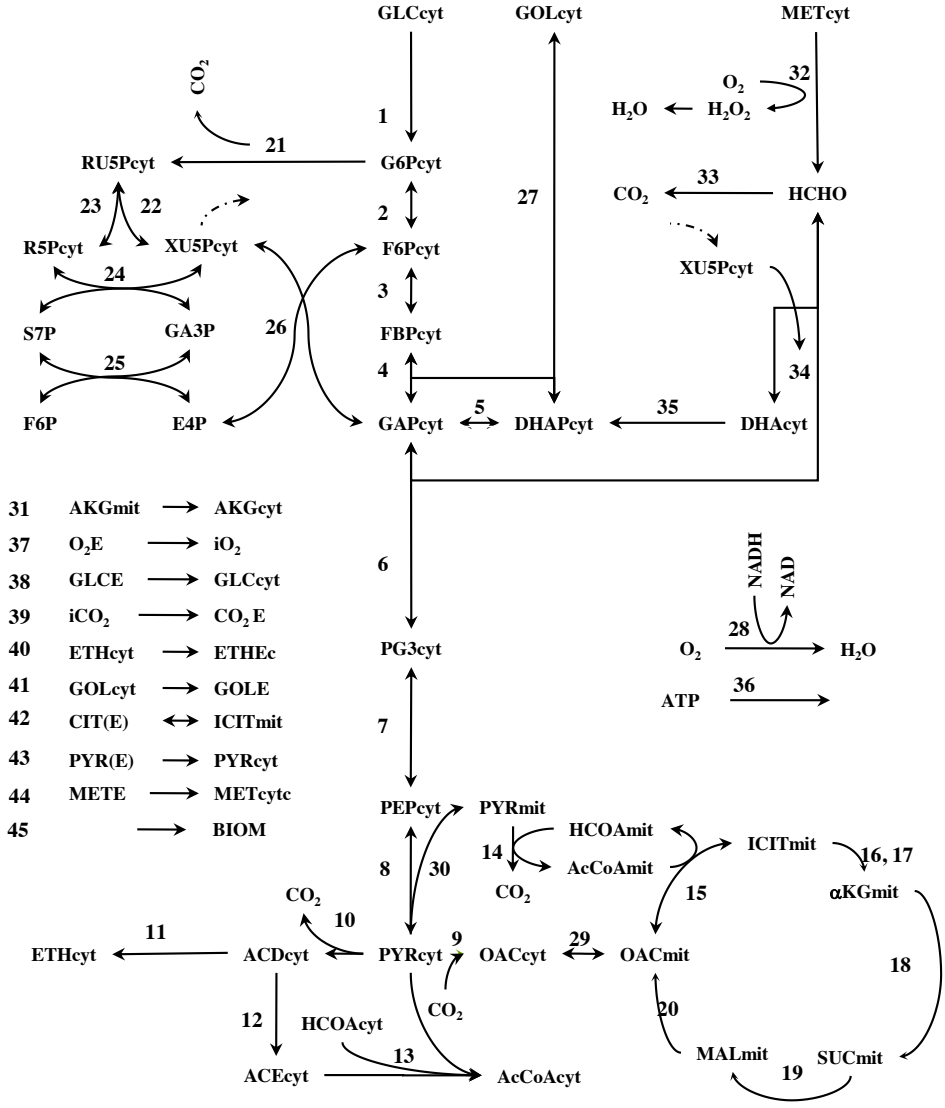


Figure 1: Summarized representation of the metabolic network of *P. pastoris*, representing the central carbon metabolism of the yeast during growth on glucose, glycerol and methanol. For the purpose of clarity, the biomass equation is not represented in the figure. Please refer to the stoichiometric matrix for details about each reaction and the involved metabolites.

At this point, there is a paramount comment that is in due. Batch effects, which are defined as systematic non-biological variation between groups of samples (or batches) due to experimental artifacts [50, 51, 52, 53], can be present in data collected from different cultures. In case that replicates of the same scenario are collected (*i.e.* same strain and same quantities of initial substrates) and the presence of batch effects is statistically confirmed, this artificial variation must be removed. Otherwise, the bias introduced by the non-biological nature of this kind of effects may confound true biological differences [52], affecting the results of statistical analysis. In this study, the scenarios within a single strain of *P. pastoris* have different initial substrate quantities (see Figure 2). Hence, the variation observed across scenarios can be due to these different initial conditions, which were applied with the aim of obtaining different flux values. This

fact jointly with the scarcity of information about the experimentation conditions disable the possibility to straightforwardly confirm actual batch effects in data.

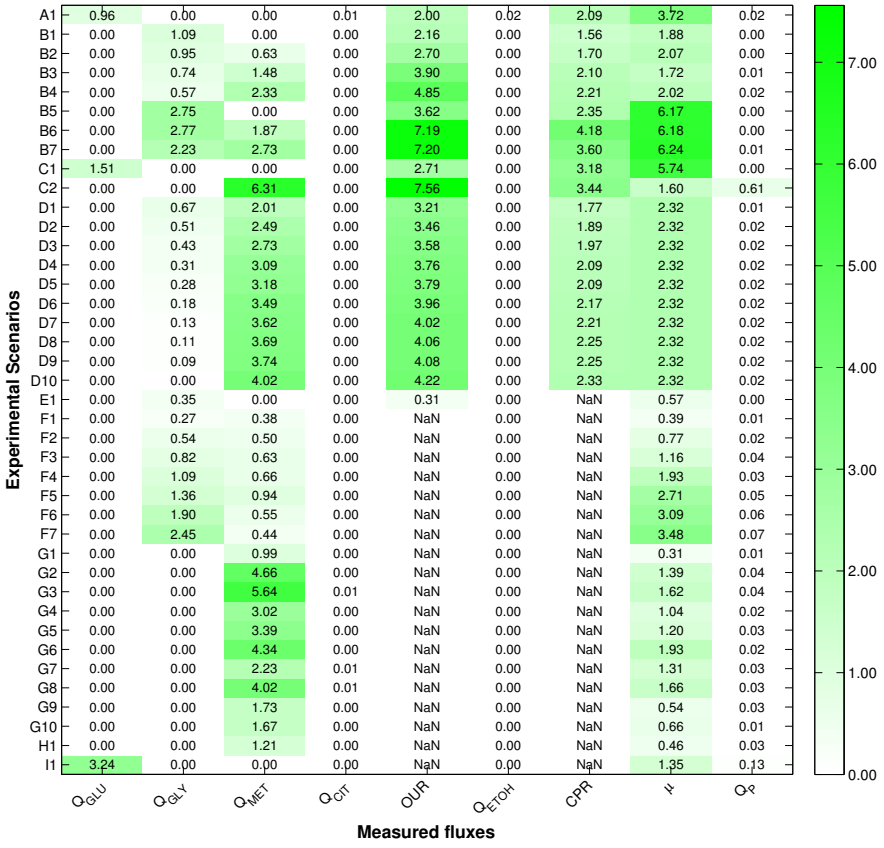


Figure 2: Set of 40 experimental scenarios corresponding to *P. pastoris* chemostat cultures grown on glucose, glycerol and methanol mixtures. For each scenario, the values of measured fluxes belonging to substrate and product specific consumption and production are shown. The substrates are glucose (Q_{GLU}), glycerol (Q_{GLYC}), methanol (Q_{MET}), citrate (Q_{CIT}) and oxygen (OUR). The products are ethanol (Q_{ETOH}), carbon dioxide (CPR), biomass (μ) and protein (Q_P). Note that NaN values stand for missing measured external fluxes.

3. Methods

The grey modelling approach proposed is composed of several steps (see Figure 3). Firstly, the constraint-based model is built by transforming the network in a mathematical form (the stoichiometric matrix **S** and the flux irreversibilities are detailed in the Supplementary Information). At the same time, different experimental fermentation scenarios are collected from the literature. The combination of these two steps represents the novel grey modelling approach, detailed in the previous sections. Then, a Possibilistic consistency analysis is performed to elucidate which scenarios are not consistent with the model. On the consistent scenarios, a Monte Carlo sampling is applied to obtain a hundred different feasible solutions satisfying the proposed model. With the feasible flux solutions matrix, a Principal Component Analysis (PCA) is performed with the aim of getting an insight of the metabolic structure of the yeast. All the outliers are detected

by using two Shewhart-type control charts based on the Hotelling-T^2 and Square Prediction Error (SPE) statistics and, later on, root causes are diagnosed. Once the ouliers are isolated, PCA is computed again. This procedure is repeated until the percentage of outliers is consistent with the confidence limits (99% confidence level)). Finally, the Missing-data method for Exploratory Data Analysis (MEDA) is applied in order to attain more informative components. The theory behind these steps are described in the following subsections.

Stoichiometric modelling

To build a constraint-based model, the stoichiometric information embedded in the metabolic network (*i.e.* metabolites or cofactors involved in each reaction) must be arranged into a $I \times J$ matrix $\mathbf{S}$ (the so-called stoichiometric matrix). Rows of this matrix represent the I metabolites, columns the J metabolic reactions and each element (i, j) the stoichiometric coefficient $S_{i,j}$ of the ith metabolite in the jth reaction. A value of $S_{i,j} = -1$ indicates that the ith metabolite is consumed by the jth reaction. In contrast, a $S_{i,j} = 1$ indicates the ith metabolite is produced by the jth reaction. Finally, a value of $S_{i,j} = 0$ stands for the ith metabolite is not involved in the jth reaction.

The stoichiometrix matrix is used, in combination with the flux vector $\mathbf{v} = (v_1, ..., v_J)$, the intracellular metabolites concentration $\mathbf{c} = (c_1, ..., c_I)$ and the specific growth rate of the cell μ, to represent the mass balances through the metabolic network. The ordinary differential equation describing this process is as follows:

$$\frac{d\mathbf{c}}{dt} = \mathbf{S} \cdot \mathbf{v} - \mu \cdot \mathbf{c} \tag{1}$$

This equation is called the dynamic mass balance equation, and describes the evolution of the concentration of each metabolite over time [11]. In stoichiometric modelling, the dynamic intracellular behaviour is disregarded on the basis assumption of pseudosteady state for the internal metabolites [8]. This assumption is supported by the observation that intracellular dynamics are much faster than extracellular dynamics. Therefore, it is sensible to assume that these compounds reach the steady state instantaneously and, hence, its transient behavior can be omitted. In addition, the dilution term $\mu \cdot \mathbf{c}$ is also discarded because it is generally much smaller than the fluxes affecting the same metabolite. Under these considerations, the general equation can be expressed as:

$$\mathbf{S} \cdot \mathbf{v} = \mathbf{0} \tag{2}$$

This equation constrains the J-dimensional space of feasible solutions. An extra constraint is added, assuming that some of the fluxes of the metabolic network flow only in one direction:

$$\mathbf{D} \cdot \mathbf{v} \geq \mathbf{0} \tag{3}$$

where $\mathbf{D}$ is a $J \times J$ diagonal matrix with binary values: 1 for the irreversible fluxes and 0 for the reversible ones.

Finally, a maximum value for each flux value is computed:

$$v_j \leq v_{j,max} \quad \forall j \in 1, ..., J \tag{4}$$

The combination of the constraints imposed by Equations 2-4 define a space (a bounded convex cone) of feasible steady-state flux distributions: only flux vectors that fulfill Equation 2-4 are considered valid cellular states. In this way, Equations 2-4 define our model of Pichia pastoris, following a constraint-based modelling approach [9, 10, 11].

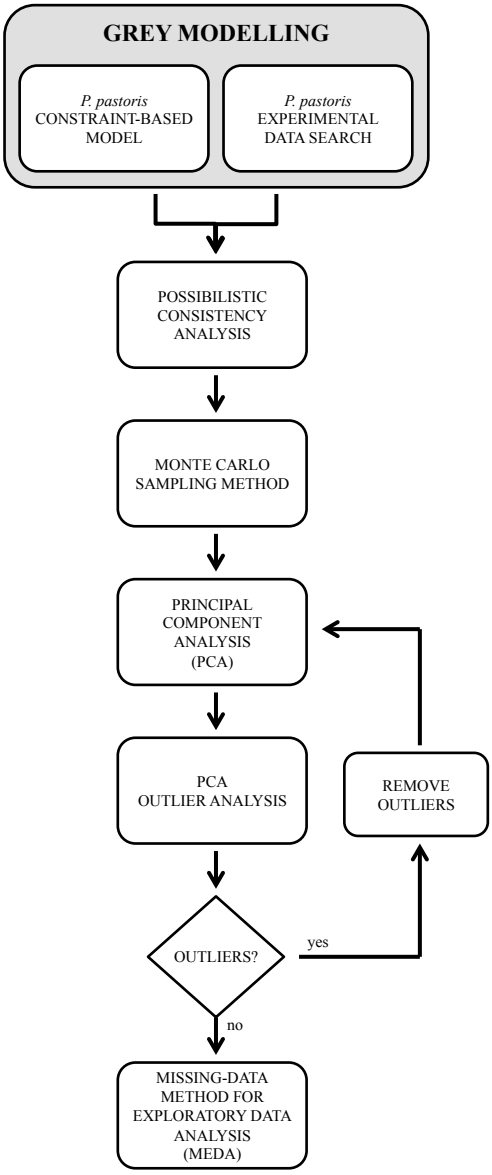


Figure 3: Flow diagram of the grey modelling approach of *P. pastoris*.

Possibilistic consistency analysis

The simplest consistency analysis could be performed by checking that the flux states shown by cells fulfill the constraints imposed by the model (see Equations 2-4) [3]. However, this simple approach would be impractical because measurements are imprecise and do not exactly satisfy the constraints. Such difficulty is overcome by taking into account uncertainty as follows:

$$w = v_m + e \tag{5}$$

where e represents the deviation error between an actual flux v_m and its measured value w.

The consistency analysis can be also formulated as a possibilistic constraint satisfaction problem [43]. The basic idea is that a flux vector fulfilling Equations 2 and 3, and compatible with the measurements will be considered as "possible", otherwise as "impossible". This can be refined to cope with measurements errors by introducing the notion of "degree of possibility" [44].

This degree of possibility provides an indication of the consistency between the model and the measurements. A possibility equal to one must be interpreted as complete agreement between the model and the original measurements. Lower values of possibility imply that certain error in the measurements is needed to find a flux vector fulfilling the model constraints. For further details readers are referred to [3, 43].

The main formulation of Possibilistic consistency analysis is summarized in this section.

Model and measurements constraints. Firstly we consider the constraints conforming the model (Equation 2-4). Then measures of (some) extracellular fluxes are incorporated as additional constraints (Equation 6):

$$\begin{cases} \mathbf{v}_m = \mathbf{w} + \boldsymbol{\varepsilon}_1 - \boldsymbol{\mu}_1 + \boldsymbol{\varepsilon}_2 - \boldsymbol{\mu}_2 \\ \boldsymbol{\varepsilon}_1, \boldsymbol{\mu}_1, \boldsymbol{\varepsilon}_2, \boldsymbol{\mu}_2 \geq \mathbf{0} \\ \boldsymbol{\varepsilon}_2 \leq \boldsymbol{\varepsilon}_{2,max} \\ \boldsymbol{\mu}_2 \leq \boldsymbol{\mu}_{2,max} \end{cases} \tag{6}$$

where vector $\mathbf{v}_m$ represents the actual metabolite concentrations and $\mathbf{w}$ the measured values, which differ due to errors and imprecision (uncertainty). This uncertainty is represented by the vectors of slack variables $\boldsymbol{\varepsilon}$'s and $\boldsymbol{\mu}$'s.

Possibility. Let us denote each candidate solution of Equation 6 as $\delta = \{\mathbf{v}, \mathbf{w}, \boldsymbol{\varepsilon}, \boldsymbol{\mu}\}$ in Δ. The basic building block of possibility theory is a user-defined possibility distribution $\pi(\delta) : \Delta \rightarrow [0, 1]$. This function defines the possibility of each solution δ in Δ, ranging between impossible ($\pi = 0$) and fully possible ($\pi = 1$). Among different possible choices, a simple -yet sensible- way to define possibility is using a linear cost index such as Equation :

$$J(\delta) = \boldsymbol{\alpha}^T \cdot \boldsymbol{\varepsilon}_1 + \boldsymbol{\beta}^T \boldsymbol{\mu}_1 \tag{7}$$

and define the possibility of each solution δ as follows:

$$\pi(\delta) = e^{-J(\delta)}, \quad \delta \in \Delta \tag{8}$$

where $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ are row vectors of user-defined, sensor accuracy coefficients.

The interpretation of Equations 6-8 may be: $\mathbf{v} = \mathbf{w}$ is fully possible; the more $\mathbf{v}$ and $\mathbf{w}$ differ, the less possible such situation is

Representing uncertainty. Two pairs of vectors of slack variables have been chosen to represent the uncertainty of each measurement: $\boldsymbol{\varepsilon}_2$ and $\boldsymbol{\mu}_2$ define an interval of fully possible values, and $\boldsymbol{\varepsilon}_1$ and $\boldsymbol{\mu}_1$ penalise values out of it (with weights $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$). This is achieved choosing two vectors of bounds. Hence, in all computations the uncertainty of each measurement has been represented as follows:

(i) Full possibility ($\pi = 1$) is assigned to values with less than $\pm$ 5% of deviation.

(ii) Larger deviations are penalised, so values with a deviation equal to $\pm$ 20% have a possibility of $\pi = 0.1$, and those with a deviation equal to $\pm$ 10% have a $\pi \approx 0.5$.

(iii) Uncertainty is considered as symmetric, and thus $\boldsymbol{\alpha} = \boldsymbol{\beta}$.

This is achieved choosing bounds $\boldsymbol{\varepsilon}_{2,max}$ and $\boldsymbol{\mu}_{2,max}$ and weights $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ for each measurement: (i) implies that $\boldsymbol{\varepsilon}_{2,max} = \boldsymbol{\mu}_{2,max} = 0.05 \cdot \mathbf{w}$, and (ii) defines $\boldsymbol{\alpha}$, noticing that $0.2 \cdot \mathbf{w} = \boldsymbol{\mu}_{1,20\%} + \boldsymbol{\mu}_{2,max}$, then $\boldsymbol{\alpha} = -log(0.1)/(0.2 - 0.05)/\mathbf{w}$.

Possibilisitic consistency evaluation. This method can be applied to evaluate the degree of consistency between a given model and a set of experimental measurements. Notice that the most possible solution of the constraint-satisfaction problem is the maximum possibility (minimum-cost) solution, which can be obtained solving a linear programming problem (LP):

$$J^{min} = \min_{\boldsymbol{\varepsilon},\boldsymbol{\mu},\mathbf{v}} J \tag{9}$$

subject to Equations 2-4 and the experimental measurements. This solution has an associated degree of possibility:

$$\pi^{mp} = e^{-J^{min}} \tag{10}$$

This value, π^{mp} in [0,1], grades the consistency between model and measurements. A possibility equal to one must be interpreted as complete agreement, while lower values imply that there is some error in the measurements, the model or both, which severity depends on how the uncertainty has been defined (see above). More details on Possibilistic consistency analysis are given in a previous work, where the model of P. pastoris was validated [3, 49].

Monte Carlo sampling method

Metabolic Flux Analysis was designed with the aim of obtaining the flux values of all reactions based on the known fluxes, typically extracellular, which are easier to measure. Assuming the J_1 measured fluxes of the $J = J_1 + J_2$ fluxes of the metabolic network, the J_2 unmeasured fluxes can be derived from the general equation of the stoichiometric modelling (see Equation 2):

$$\mathbf{S}_{J_1} \cdot \mathbf{v}_{J_1} = -\mathbf{S}_{J_2} \cdot \mathbf{v}_{J_2} \tag{11}$$

where $\mathbf{S}_{J_1}$ and $\mathbf{v}_{J_1}$ involves the measured fluxes (the external ones), and $\mathbf{S}_{J_2}$ and $\mathbf{v}_{J_2}$ are related to the unmeasured fluxes (the internal ones). The problem with this formulation is that the number of internal fluxes often remain high compared to the number of external fluxes. Thus, the system shown in Equation 11 is undetermined, *i.e.* there are different flux distributions compatible with the known flux values.

In this context, Monte Carlo sampling methods can be used to produce feasible flux distributions across the cell [19, 20, 21, 54, 55]. This way, the available experimental data and the first-principles knowledge captured by the model are coupled together, providing a new richer data-set amenable to further analysis with a statistical multivariate projection method. To randomly generate possible values for the unmeasured fluxes (internal fluxes) for each cultivation, stoichiometry (see Equation 2), irreversibility (see Equation 3) and measured fluxes (see Equation 4 and experimental data on Figure 2) are taken into account.

In order to deal with experimental errors, external fluxes are allowed to vary within a defined range of values centered on the original measured value. The upper (lower) bound of this range is the sum (subtraction) of the measured value and the maximum value between 0.001 and the 10% of the measured value:

$$(LB, UB)_j = (v_j - max(0.001, 0.1 \times v_j), v_j + max(0.001, 0.1 \times v_j)) \quad \forall j \in 1, ..., J \tag{12}$$

where v_j is a measured flux, and LB and UB are the lower and upper bounds for the Monte Carlo sampling method.

At this point, it is worth commenting that the feasible solutions for each scenario are obtained by sampling within the *slice* of the cone defined by Equations 2-4 and the experimental data, *i.e.* the measured fluxes reduce the feasible solution space from the initial cone, which is bounded only by the constraint-based model, to the portion of it fulfilling these specific experimental measurements. The complete procedure can be visualized in Figure 4.

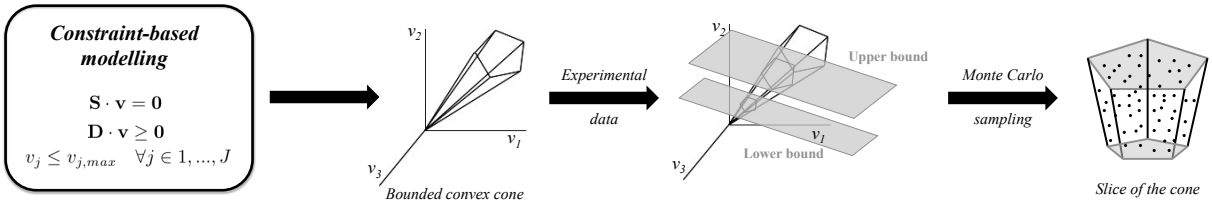


Figure 4: Constraint-based modelling and allowed flux states sampling. The convex cone is obtained by Equations 2-4. The experimental measurements constrain the cone through Equation 12. Finally, the sampling is performed on the resulting *slice* of the cone.

Principal Component Analysis

The aim of Principal Component Analysis (PCA) is to find the subspace in the space of the variables where data mostly vary [56]. The original variables, commonly correlated, are linearly transformed into a lower number of uncorrelated variables (the so-called principal components, PCs). PCA follows the expression:

$$\mathbf{X} = \mathbf{T}_A \cdot \mathbf{P}_A^t + \mathbf{E}_A \tag{13}$$

where $\mathbf{X}$ is a $N \times M$ matrix of data, $\mathbf{T}_A$ is the $N \times A$ scores matrix containing the projection of the objects in the A PCs subspace, $\mathbf{P}_A$ is the $M \times A$ loadings matrix containing the linear combination of the variables represented in each of the PCs, and $\mathbf{E}_A$ is the $N \times M$ matrix of residuals.

As a previous step of PCA, the data matrix is autoscaled, i.e. each variable (flux) is centered and divided by its standard deviation, making all variables have a variance equal to 1. In the present work, since the components obtained by PCA are linear combinations of different fluxes, the more positive the coefficient of a flux is, the more positive correlated is with this particular component, in the sense that the flux is higher than its mean value in this component. As well, the more negative its coefficient is, the more negative correlated is with this component, in the sense that the flux is lower than its mean value in this component. In other words, fluxes with positive coefficients in a component are overused, and fluxes with negative coefficients are underused.

PCA outlier detection

Square Prediction Error (SPE) and Hotelling-T^2 are two statistics widely used to detect outliers on a given data. SPE is the orthogonal distance of a particular object to the A-dimensional subspace of latent variables defined by PCA. It is expressed as:

$$SPE_n = \mathbf{e}_n^t \cdot \mathbf{e}_n \quad \forall n \in 1, ..., N \tag{14}$$

where $\mathbf{e}_n$ is the nth row of the residual matrix $\mathbf{E} = \mathbf{X} - \mathbf{T}_A\mathbf{P}_A^t$. By taking the eigenvalues of the covariance matrix of the residual matrix $(\lambda_{A+1}, \ldots, \lambda_M)$, the control limit of the SPE [57] is computed as follows:

$$SPE_\alpha = \theta_1 \left[\frac{z_\alpha \sqrt{2\theta_2 h_0^2}}{\theta_1} + 1 + \frac{\theta_2 h_0 (h_0 - 1)}{\theta_1^2} \right]^{1/h_0} \tag{15}$$

where $\theta_m = \sum_{j=A+1}^{M} (\lambda_j)^m$, $h_0 = 1 - \frac{2\theta_1\theta_3}{3\theta_2^2}$ and z_α is the $100 \times (1-\alpha)$ percentile of a standard Normal distribution.

Hotelling-T^2 is a statistic based on the Mahalanobis distance [58]. This statistic is used in multivariate monitoring to compute the distance between one object and model's centre according to the covariance structure [59]. When the data is centered (the mean of each column of $\mathbf{X}$ is equal to zero), the distance between an observation $\mathbf{x}_n$ (a row from the data matrix $\mathbf{X}$) and the centre of the original M-dimensional variable space is:

$$\chi^2_n = \mathbf{x}^t_n \cdot \mathbf{\Sigma}^{-1} \cdot \mathbf{x}_n \quad \forall n \in 1, ..., N \tag{16}$$

where $\mathbf{\Sigma}$ is the real covariance matrix of the original M-dimensional variable space, and χ^2_n follows a χ^2 distribution with M degrees of freedom. In practice, the mean and the covariance matrix are estimated by the data matrix $\mathbf{X}$ as $\mathbf{S} = \mathbf{X}^t\mathbf{X}/(N-1)$. So the approximation to the Mahalanobis distance is the Hotelling-T^2:

$$T^2_n = \mathbf{x}^t_n \cdot \mathbf{S}^{-1} \cdot \mathbf{x}_n \quad \forall n \in 1, ..., N \tag{17}$$

where $\mathbf{x}_n$ is the nth row of the data matrix $\mathbf{X}$, corresponding to a concrete object. The control limit for the Hotelling-T^2 [60] is computed as :

$$T^2_\alpha = \frac{(N^2-1)A}{N(N-A)} F_\alpha(A, N-A) \tag{18}$$

where A is the number of PCs of the model, and $F_\alpha(A, N-A)$ is the $100 \times (1-\alpha)$ percentile of a Snedecor's F distribution with $(A, N-A)$ degrees of freedom.

MEDA

The Missing-data methods for Exploratory Data Analysis (MEDA) [61] can be seen as a substitute of rotation methods with better properties. First of all, it is more accurate than rotation methods in the detection of relations between pairs of variables. Also, it is robust to the overestimation of the number of PCs and it does not depend on the normalization of the loadings.

Once the PCA has been performed, the MEDA approach consists of the following steps for each variable m ($m = 1, \ldots, M$):

1. Construct matrix $\tilde{\mathbf{X}}_m$, which is a $N \times M$ matrix full with zeros except in the mth column where it contains the mth column of matrix $\mathbf{X}$, $\mathbf{x_m}$.

$$\tilde{\mathbf{X}}_m = [\mathbf{0} \ldots \mathbf{0} \quad \mathbf{x_m} \quad \mathbf{0} \ldots \mathbf{0}] \tag{19}$$

2. Estimate the scores from $\tilde{\mathbf{X}}_m$ using a missing-data method. In this case, the Known-Data Regression (KDR) method is applied, which has been proved to be statistically superior to other missing data imputation techniques in [62].

$$\hat{\mathbf{T}}_A = MD(\tilde{\mathbf{X}}_m) \tag{20}$$

3. Estimate the reconstruction of the original measurements with A latent variables and compute the estimation error

$$\hat{\mathbf{X}}_A = \hat{\mathbf{T}}_A \cdot \mathbf{P}^t_A \tag{21}$$

$$\hat{\mathbf{E}}_A = \mathbf{X} - \hat{\mathbf{X}}_A \tag{22}$$

where $\mathbf{P}$ is the estimated loadings matrix from $\mathbf{X}$ (the complete $N \times M$ matrix of data), $\hat{\mathbf{X}}_A$ is the estimation matrix and $\hat{\mathbf{E}}_A$ the estimation error matrix.

4. Compute an index of goodness of prediction [63] in all columns but the mth one

$$Q^2_{A,(m,l)} = 1 - \frac{\sum_{n=1}^{N} (\hat{E}_{A,(n,l)})^2}{\sum_{n=1}^{N} (X_{n,l})^2}, \quad \forall l \neq m \tag{23}$$

where $X_{n,l}$ is the element located at the nth row and the lth column of $\mathbf{X}$, and $\hat{E}_{A,(n,l)}$ is its estimation error. The closer $Q^2_{A,(m,l)}$ is to 1, the more related variables m and l are.

Once the values of $Q^2_{A,(m,l)}$ for all possible combinations of m and l are computed, a matrix $\mathbf{Q}^2_A$ can be constructed so that $Q^2_{A,(m,l)}$ is located at row m and column l. This matrix is similar in nature to the element-wise squared correlation matrix. Structural relations between variables are detected as high values in $\mathbf{Q}^2_A$, but the direct/inverse pair-wise relation is not represented on the matrix because of the squared values. To avoid obvious relations, the values of principal diagonal of $\mathbf{Q}^2_A$ matrices are set to zero. When the number of variables is large, matrix $\mathbf{Q}^2_A$ can be shown as a grey map to improve interpretability.

The $\mathbf{Q}^2_A$ matrices have been built in a cumulative way, *i.e.* they have the variability of the first A PCs. These matrices can also be constructed by taking the information of a single PC. For this purpose, the method previously detailed has to be changed in Equations 20-23. The new equations have to consider only the a-th component for estimation. Finally, this kind of MEDA matrices, which have been used in this work, are written as $\mathbf{Q}^2_{(a)}$, where $a = 1, \ldots, A$.

Software

All methods commented in this Section have been computed in Matlab environment. The Monte Carlo sampling method has been applied using the COBRA toolbox [64]. PCA has been performed on MATLAB's Statistical Toolbox. Finally, MEDA has been performed using Explanatory Data Analysis Toolbox [65].

4. Results and Discussion

In this section, the grey modelling approach proposed is applied to the methylotrophic yeast *P. pastoris* to discover patterns of heterologous protein production and correlate intra and extracellular reactions in order to understand how the internal state of the cells determines their observed behavior.

Possibilistic consistency analysis

The different scenarios collected from the literature are combined with the proposed model in order to validate which ones are consistent and which ones are not. From each one of the 40 scenarios, the flux values through the external reactions, which are different depending on the initial conditions of each experiment, are validated against the stoichiometric modelling of the *P. pastoris*. As explained in Section 3, the most possible solution for each scenario (*i.e.* experimental dataset) is computed to perform a Possibilistic consistency analysis. The corresponding possibility values (π) are shown in Table 1. The majority of datasets are highly consistent with the model (65% are fully possible, and 87% have a possibility higher than 0.5). There are, however, 4 out of 40 datasets with a possibility lower than 0.25 (*i.e.* a possibility that is equivalent to an error of 14% in one measurement, or to an error of 8% in three measurements). These scenarios (B3, B4, C2, and E1) are not fully consistent with the model. The inconsistency can be due to (a) limitations of the model, which may be unable to capture phenomena ocurring in those experiments, (b) larger errors than expected in the data measured in those scenarios, or (c) the two previous reasons acting simultaneously. For this reason, we decided to remove these scenarios (B3, B4, C2, and E1) so they are not considered in the following analysis.

Monte Carlo sampling

The previous analysis concludes that 36 out of the 40 scenarios are consistent with the model. However, only the external fluxes of each solution have been measured. Due to the complexity of measuring the internal fluxes, the Monte Carlo sampling method is proposed to simulate different possible flux solutions, consistent with the proposed model and the measured subset of fluxes, in order to get enough complete flux solutions to be analysed.

Once the sampling has been performed, a feasible flux solution matrix $\mathbf{X}$ is built. $\mathbf{X}$ has the complete 3600 sampled flux solutions in its rows (36 scenarios $\times$ 100 samples) and the corresponding 45 flux values and the protein production for each scenario in its columns (see Additional file 2).

Scenario	**Group**	π
A1	*glucose*	1,000
B1	*glycerol*	1,000
B2	*glycerol + methanol*	0,739
B3	*glycerol + methanol*	0,246(*)
B4	*glycerol + methanol*	0,082(*)
B5	*glycerol*	1,000
B6	*glycerol + methanol*	0,819
B7	*glycerol + methanol*	0,319
C1	*glucose*	0,658
C2	*methanol*	0,052(*)
D1	*glycerol + methanol*	1,000
D2	*glycerol + methanol*	1,000
D3	*glycerol + methanol*	1,000
D4	*glycerol + methanol*	1,000
D5	*glycerol + methanol*	1,000
D6	*glycerol + methanol*	0,908
D7	*glycerol + methanol*	0,709
D8	*glycerol + methanol*	0,637
D9	*glycerol + methanol*	0,614
D10	*methanol*	0,500
E1	*glycerol*	0,065(*)
F1	*glycerol + methanol*	1,000
F2	*glycerol + methanol*	1,000
F3	*glycerol + methanol*	1,000
F4	*glycerol + methanol*	1,000
F5	*glycerol + methanol*	1,000
F6	*glycerol + methanol*	1,000
F7	*glycerol + methanol*	1,000
G1	*methanol*	1,000
G2	*methanol*	1,000
G3	*methanol*	1,000
G4	*methanol*	1,000
G5	*methanol*	1,000
G6	*methanol*	1,000
G7	*methanol*	1,000
G8	*methanol*	1,000
G9	*methanol*	1,000
G10	*methanol*	1,000
H1	*methanol*	1,000
I1	*glucose*	1,000

Table 1: Possibility values (π) for each scenario. Those scenarios that are not consistent (i.e. $\pi < 0.25$) with the constrained-based model are signaled with (*).

Principal Component Analysis

A PCA is performed to the feasible flux solutions matrix $\mathbf{X}$ to obtain a low number of principal components (PCs) explaining a high variance percentage of the complete data set. These PCs are linear combinations of the actual 46 variables (45 flux values of the reactions and the protein production).

After the PCA has been fitted, explaining 94.3% of total variance with five PCs, the outlier analysis is applied to the scores and the residuals of the different scenarios. The results on the statistic SPE show that scenario *C1*, classified on group *glucose* widely exceed the control limits. Thus, the hundred observations generated by the Monte Carlo sampling for this scenario are knocked-out. Afterwards, a new PCA is fitted.

The results with the second analysis are that the first five PCs capture 95.9% of total variance in data: 42.4% for the first component, 24.2% for the second one, 19.7% for the third one, 7.0% for the fourth, and 2.5% for the last one. In these results no outliers are detected.

Enhancement of model interpretation

The MEDA is applied to the first five components obtained by the PCA, and the $\mathbf{Q}^2_{(a)}$, $a = 1, \ldots, 5$ matrices are obtained. Looking at the values in each matrix, the first three PCs are sufficient to explain the behaviour of the yeast, which capture 86,3% variance in data. Fourth and fifth PCs are classified as noise. The first three MEDA matrices can be seen in Figure 5.

If analysed from a biological standpoint, the first principal component relates protein production rate to reactions 5-8 (glycolysis), 14-16 and 18 (TCA cycle), 19-20, 28, 30, 36 and 37. In Figure 5a these reactions are rounded by the solid line rectangle. It can be seen that this relations are indeed strongly correlated, having $Q^2_{(1),(m,l)}$ coefficients close to 1. As can be seen in the stoichiometric matrix, each of these groups is directly connected to NADH and ATP metabolism: ATP is formed in reactions 6, 8, 18 and 28, whereas NADH is formed in reactions 6, 14, 16 and 18-20. Finally, reactions 28, 30 and 36 represent the electronic transport chain, oxygen consumption and ATP dissimilation. The first PC can be then understood as the main pathway for ATP formation and dissimilation, this is, energy generation. Interestingly, protein productivity and ATP generation have been previously related in a first-principles based approach to predict recombinant protein production [49].

The second principal component is related to the biomass growth rate, which involves reactions 9-13 (fermentative pathways), 17, 21, 29 and 41 (relations shown by dashed line rectangles in Figure 5b). Except for reaction 41, corresponding to the glycerol consumption rate, reactions 12 (around which reactions 9, 10, 11, 13 and 29 are connected), 17 and 21 share NADPH (either mitochondrial or cytosolic) production, which is, in fact, one of the major contributing precursors to biomass formation. It is worth noting that reaction 17 (corresponding to NADPH-requiring form) and not 16 (corresponding to the isoenzyme NADH-requiring) is identified.

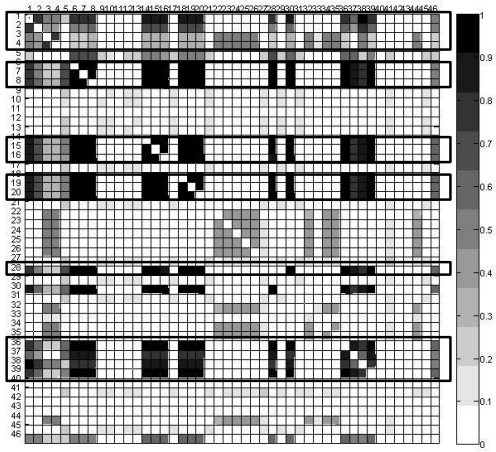
Finally, the third principal component relates methanol consumption rate to the pentose phosphate pathway, strongly connected by reaction 34 (reactions correlated are rounded by dotted rectangles in Figure 5c). Reactions 3-4, 22-26, 32, 35 and 44 are also related with this component.

The first three principal pathways are depicted in Figure 6. In this way, the reactions involved by the three first principal components seem to pinpoint specific metabolic indicators (cofactors NADH, NADPH and ATP) and their relation with protein, biomass and substrate (glycerol and methanol) consumption.

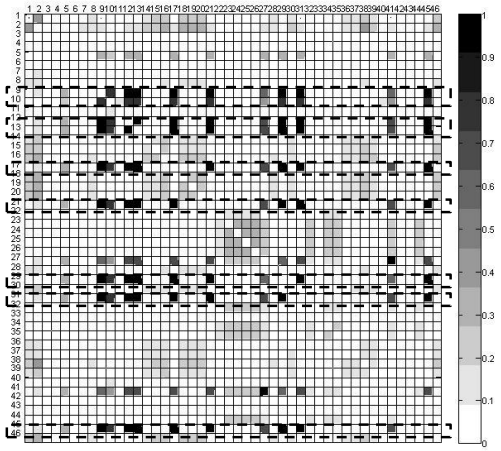
It is worth pointing out that the fit of a PCA model on the available experimental data is not feasible due to two main reasons: i) only seven out of nine external fluxes are measured for all scenarios under study, of which three have zero values mostly (see Figure 2), ii) the flux distributions across the metabolic network cannot be represented since no internal fluxes are considered. Actually, a PCA does not clearly relate substrates consumption to biomass and protein production, so this model is not meaningful (results not shown).

5. Conclusions

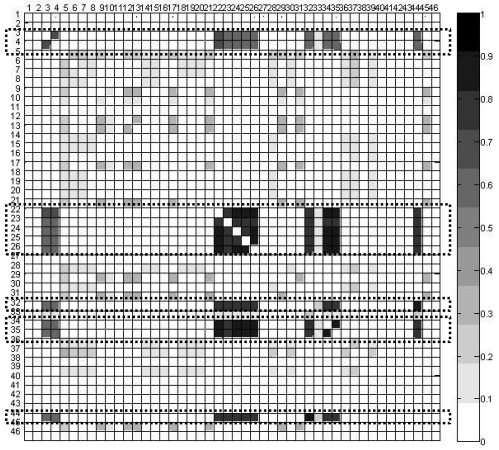
In this paper, a grey modelling strategy that combines data-driven and knowledge-based information at different scales is presented to analyse the behaviour of the methylotrophic yeast *P. pastoris*. This strategy



(a) $\mathbf{Q}^2_{(1)}$



(b) $\mathbf{Q}^2_{(2)}$



(c) $\mathbf{Q}^2_{(3)}$

Figure 5: MEDAs plots for the first (a), second (b) and third (c) PC. Solid line rectangles marks reactions related to the first PC, dashed line rectangles round reactions associated to the second PC and, finally, reactions related to the third PC are rounded by dotted line rectangles.

is composed of five main steps. Firstly, the available flux measurements, mainly external, are coupled with a model-based estimation of the unmeasured fluxes, mainly internal. Secondly, a possibilistic analysis is applied to check the consistency between the constraint-based model and data. Thirdly, a Monte Carlo sampling is performed to produce feasible flux solutions for the microbial system under study. As a result, a large solution data set -a mixture of experimental data, data-based estimations, and variability resulting from uncertainty- is obtained. Fourthly, a PCA model is fitted on the sampled data to reveal its internal biochemical structure. Finally, MEDA analysis is performed to enhance the interpretation of the Principal Components (PCs) derived from the PCA analysis.

The grey modelling of the methylotrophic yeast *P. pastoris* yielded three meaningful PCs from the biological point of view, which are sufficient to explain most of the variance of the sampled data. The first

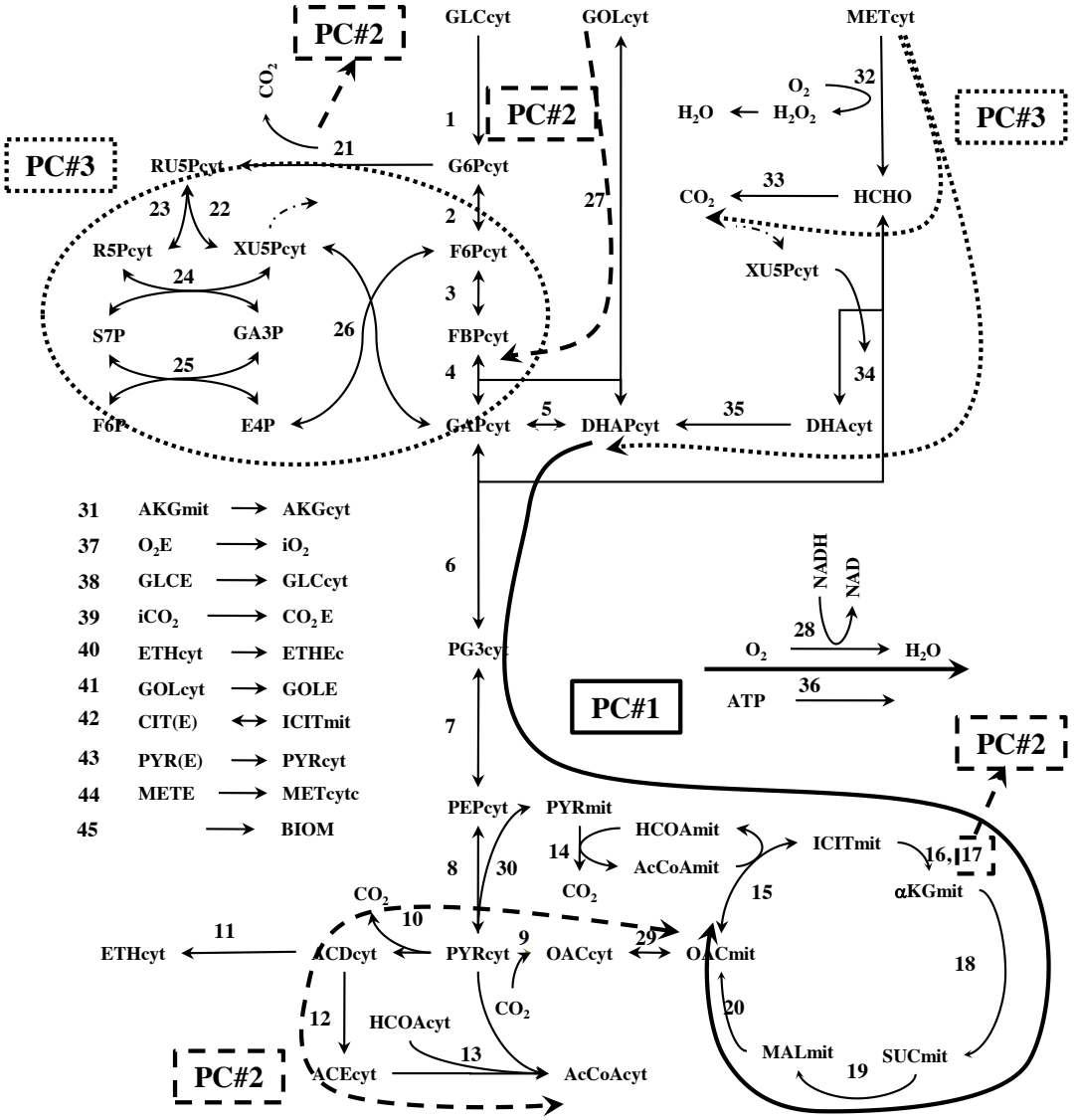


Figure 6: The first three PCs represent the main metabolic pathways through the yeast *P. pastoris*.

PC is related to protein productivity, the second one corresponds to the biomass growth rate, and the third one represents methanol consumption rate. Note that the experimental data taken from literature do not describe the whole space of behaviours that *P. pastoris* could exhibit. In order to explore all the feasible solution space, future work should address this issue in two ways: (a) a generalisation: incorporating more datasets to explore in a wider scope, and (b) a particularization: perform a similar study only with similar datasets (*e.g.* mixed cultures of *glycerol* and *methanol*), with the aim of getting a deeper insight in their differences and their impact of the process performance.

An important benefit of the grey modelling and analysis approach presented in this paper is its scalability. New knowledge, *e.g.* metabolomics or gene regulation, can be incorporated in the form of extra constraints, new flux data can be added via new scenarios, and other data pieces -not fluxes- could be incorporated directly into the data matrix before performing the multivariate methods. This joint with its capability to describe the most important biochemical processes makes this strategy promising for the design of real-time

monitoring systems.

6. Acknowledgements

Research in this study was partially supported by the Spanish Ministry of Science and Innovation and FEDER funds from the European Union through grants DPI2011-28112-C04-01 and DPI2011-28112-C04-02. The authors are also grateful to Biopolis SL for supporting this research. We also gratefully acknowledge Associate Professor José Camacho for providing the Exploratory Data Analysis Toolbox.

References

[1] F. Llaneras, Interval and possibilistic methods for constraint-based metabolic models, Master's thesis, Universidad Politcnica de Valencia (2010).
[2] A. Kayser, J. Weber, V. Hecht, U. Rinas, Metabolic flux analysis of escherichia coli in glucose-limited continuous culture. i. growth-rate-dependent metabolic efficiency at steady state, Microbiology 151 (3) (2005) 693–706.
[3] M. Tortajada, F. Llaneras, J. Picó, Validation of a constraint-based model of pichia pastoris metabolism under data scarcity, BMC Systems Biology 4 (115) (2010) 1–11.
[4] T. Benyamini, O. Folger, E. Ruppin, T. Shlomi, Flux balance analysis accounting for metabolite dilution, Genome Biology 11 (4) (2010) 1–9.
[5] M. Covert, E. Knight, J. Reed, M. Herrgard, B. Palsson, Integrating high-throughput and computational data elucidates bacterial networks, Nature 429 (2004) 92–96.
[6] M. W. Covert, C. H. Schilling, B. Palsson, Regulation of gene expression in flux balance models of metabolism, Journal of Theoretical Biology 213 (1) (2001) 73 – 88.
[7] M. Åkesson, J. Förster, J. Nielsen, Integration of gene expression data into genome-scale metabolic models, Metabolic Engineering 6 (4) (2004) 285 – 293.
[8] G. N. Stephanopulos, A. A. Aristidou, J. Nielsen, Metabolic engineering: principles and methodologies, Academic Press, San Diego, 1998.
[9] B. Palsson, Systems Biology: Properties of Reconstructed Networkss, Cambridge University Press, New York, USA, 2006.
[10] N. D. Price, J. A. Papin, C. H. Schilling, B. O. Palsson, Genome-scale microbial in silico models: the constraints-based approach, Trends in Biotechnology 21 (4) (2003) 162 – 169.
[11] F. Llaneras, J. Picó, Stoichiometric modelling of cell metabolism, Journal of Bioscience and Bioengineering 105 (1) (2008) 1–11.
[12] M. D. Mesarovic, S. N. Sreenath, J. D. Keene, Search for organising principles: understanding in systems biology, Systems Biology, IEE 1 (1) (2004) 19–27.
[13] K. J. Kauffman, P. Prakash, J. S. Edwards, Advances in flux balance analysis, Current Opinion in Biotechnology 14 (5) (2003) 491 – 496.
[14] S. Feyodeazevedo, B. Dahm, F. Oliveira, Hybrid modelling of biochemical processes: A comparison with the conventional approach, Computers & Chemical Engineering 21 (1997) S751–S756.
[15] H. Ramaker, E. van Sprang, S. Gurden, J. Westerhuis, A. Smilde, Improved monitoring of batch processes by incorporating external information, Journal of Process Control 12 (4) (2002) 569 – 576.
[16] Y. Takane, T. Shibayama, Principal component analysis with external information on both subjects and variables, Psychometrika 56 (1) (1991) 97–120.
[17] Y. Takane, H. Kiers, J. Leeuw, Component analysis with different sets of constraints on different dimensions, Psychometrika 60 (2) (1995) 259–280.
[18] J. E. Jackson, A User's Guide to Principal Components, Wiley Series in Probability and Statistics, 1991.
[19] B. Sariyar, S. Perk, U. Akman, A. Hortasu, Monte carlo sampling and principal component analysis of flux distributions yield topological and modular information on metabolic networks, Journal of Theoretical Biology 242 (2) (2006) 389–400.
[20] C. Barrett, M. Herrgard, B. Palsson, Decomposing complex reaction networks using random sampling, principal component analysis and basis rotation, BMC Systems Biology 3 (30) (2009) 1–8.
[21] S. Van Dien, S. Iwatani, Y. Usuda, K. Matsui, Theoretical analysis of amino acid-producing escherichia coli using a stoichiometric model and multivariate linear regression, Journal of Bioscience and Bioengineering 102 (1) (2006) 34–40.
[22] R. Bro, PARAFAC. tutorial and applications, Chemometrics and Intelligent Laboratory Systems 38 (2) (1997) 149–171.
[23] M. Verouden, R. Notebaart, J. Westerhuis, M. van der Werf, B. Teusink, A. Smilde, Multi-way analysis of flux distributions across multiple conditions, Journal of Chemometrics 23 (7-8) (2009) 406–420.
[24] S. Wold, M. Sjöström, L. Eriksson, PLS-regression: a basic tool of chemometrics, Chemometrics and Intelligent Laboratory Systems 58 (2) (2001) 109–130.
[25] N. Carinhas, V. Bernal, A. Teixeira, M. Carrondo, P. Alves, R. Oliveira, Hybrid metabolic flux analysis: Combining stoichiometric and statistical constraints to model the formation of complex recombinant products, BMC Systems Biology 5 (34) (2011) 1–13.
[26] F. Dieterle, A. Ross, G. Schlotterbeck, H. Senn, Metabolite projection analysis for fast identification of metabolites in metabonomics. application in an amiodarone study, Analytical Chemistry 78 (11) (2006) 3551–3561.

[27] K. Tondel, U. Indahl, A. Gjuvsland, J. Vik, P. Hunter, S. Omholt, H. Martens, Hierarchical cluster-based partial least squares regression (HC-PLSR) is an efficient tool for metamodelling of nonlinear dynamic models, BMC Systems Biology 5 (90) (2011) 1–17.
[28] J. Westerhuis, E. Derks, H. Hoefsloot, A. Smilde, Grey component analysis, Journal of Chemometrics 21 (10-11) (2007) 474–485.
[29] C. Ng, M. Hussain, Hybrid neural network - prior knowledge model in temperature control of a semi-batch polymerization process, Chemical Engineering and Processing: Process Intensification 43 (4) (2004) 559–570.
[30] H. Bechmann, H. Madsen, N. K. Poulsen, M. K. Nielsen, Grey box modeling of first flush and incoming wastewater at a wastewater treatment plant, Environmetrics 11 (1) (2000) 1–12.
[31] B. Sohlberg, Grey box modelling for model predictive control of a heating process, Journal of Process Control 13 (3) (2003) 225 – 238.
[32] J. M. F. ten Berge, A. K. Smilde, Non-triviality and identification of a constrained tucker3 analysis, Journal of Chemometrics 16 (12) (2002) 609–612.
[33] A. Teixeira, J. Dias, N. Carinhas, M. Sousa, J. Clemente, A. Cunha, M. von Stosch, P. Alves, M. Carrondo, R. Oliveira, Cell functional enviromics: Unravelling the function of environmental factors, BMC Systems Biology 5 (92) (2011) 1–16.
[34] M. Dragosits, J. Stadlmann, J. Albiol, K. Baumann, M. Maurer, B. Gasser, M. Sauer, F. Altmann, P. Ferrer, D. Mattanovich, The effect of temperature on the proteome of recombinant pichia pastoris, Journal of Proteome Research 8 (3) (2009) 1380–1392.
[35] A. Solà, P. Jouhten, H. Maaheimo, F. Sánchez-Ferrando, T. Szyperski, P. Ferrer, Metabolic flux profiling of pichia pastoris grown on glycerol/methanol mixtures in chemostat cultures at low and high dilution rates, Microbiology 153 (1) (2007) 281–290.
[36] A. Solà, Estudi del metabolisme central del carboni de pichia pastoris, Ph.D. thesis, Universitat Autònoma de Barcelona (2004).
[37] C. Jungo, I. Marison, U. von Stockar, Mixed feeds of glycerol and methanol can improve the performance of pichia pastoris cultures: A quantitative study based on concentration gradients in transient continuous cultures, Journal of Biotechnology 128 (4) (2007) 824–837.
[38] H. Ren, J. Yuan, K.-H. Bellgardt, Macrokinetic model for methylotrophic pichia pastoris based on stoichiometric balance, Journal of Biotechnology 106 (1) (2003) 53–68.
[39] M. C. d'Anjou, A. J. Daugulis, A rational approach to improving productivity in recombinant pichia pastoris fermentation, Biotechnology and bioengineering 72 (1) (2001) 1–11.
[40] S. Curvers, J. Linnemann, T. Klauser, C. Wandrey, R. Takors, Recombinant protein production with pichia pastoris in continuous fermentation - kinetic analysis of growth and product formation, Chemical Engineering and Technology 25 (8) (2002) 229–235.
[41] W. Zhang, C.-P. Liu, M. Inan, M. Meagher, Optimization of cell density and dilution rate in pichia pastoris continuous fermentations for production of recombinant proteins, Journal of Industrial Microbiology and Biotechnology 31 (7) (2004) 330–334.
[42] B. Schilling, J. Goodrick, N. Wan, Scale-up of a high cell-density continuous culture with pichia pastoris x-33 for the constitutive expression of rh-chitinase, Biotechnology Progress 17 (4) (2001) 629–633.
[43] F. Llaneras, A. Sala, J. Picó, A possibilistic framework for constraint-based metabolic flux analysis, BMC Systems Biology 3 (79) (2009) 1–22.
[44] L. Zadeh, Fuzzy sets as a basis for a theory of possibility, Fuzzy Sets and Systems 1 (1) (1978) 3–28.
[45] J. Camacho, Missing-data theory in the context of exploratory data analysis, Chemometrics and Intelligent Laboratory Systems 103 (1) (2010) 8–18.
[46] G. Potvin, A. Ahmad, Z. Zhang, Bioprocess engineering aspects of heterologous protein production in pichia pastoris: A review, Biochemical Engineering Journal 64 (2012) 91–105.
[47] S. Macauley-Patrick, M. Fazenda, B. McNeil, L. Harvey, Heterologous protein production using the pichia pastoris expression system, Yeast 22 (4) (2005) 249–270.
[48] O. Cos, R. Ramón, J. Montesinos, F. Valero, Operational strategies, monitoring and control of heterologous protein production in the methylotrophic yeast pichia pastoris under different promoters: A review, Microbial Cell Factories 5 (17) (2006) 1–20.
[49] M. Tortajada, F. Llaneras, D. Ramón, J. Picó, Estimation of recombinant protein production in pichia pastoris based on a constraint-based model, Journal of Process Control 22 (6) (2012) 1139–1151.
[50] M. Benito, J. Parker, Q. Du, J. Wu, D. Xiang, C. M. Perou, J. S. Marron, Adjustment of systematic microarray data biases, Bioinformatics 20 (1) (2004) 105–114.
[51] W. E. Johnson, C. Li, A. Rabinovic, Adjusting batch effects in microarray expression data using empirical bayes methods, Biostatistics 8 (1) (2007) 118–127.
[52] J. Luo, M. Schumacher, A. Scherer, D. Sanoudou, D. Megherbi, T. Davison, T. Shi, W. Tong, L. Shi, H. Hong, A comparison of batch effect removal methods for enhancement of prediction performance using maqc-ii microarray gene expression data, The Pharmacogenomics Journal (4) (2010) 278291.
[53] S. E. Reese, K. J. Archer, T. M. Therneau, E. J. Atkinson, C. M. Vachon, M. de Andrade, J.-P. A. Kocher, J. E. Eckel-Passow, A new statistic for identifying batch effects in high-throughput genomic data that uses guided principal components analysis, Bioinformaticsdoi:10.1093/bioinformatics/btt480.
[54] F. Hadlich, K. Nöh, W. Wiechert, Determination of flux directions by thermodynamic network analysis: Computing informative metabolite pools, Mathematics and Computers in Simulation 82 (3) (2011) 460–470.
[55] D. Machado, R. Costa, E. Ferreira, I. Rocha, B. Tidor, Exploring the gap between dynamic and constraint-based models

of metabolism, Metabolic Engineering 14 (2) (2012) 112–119.
[56] J. Camacho, J. Picó, A. Ferrer, Data understanding with PCA: structural and variance information plots, Chemometrics and Intelligent Laboratory Systems 100 (1) (2010) 48–56.
[57] J. Jackson, G. S. Mudholkar, Control procedures for residuals associated with principal component analysis., Technometrics 21 (3) (1979) 341–349.
[58] P. C. Mahalanobis, On the generalised distance in statistics, Proceedings of the National Institute of Sciences of India 2 (1) (1936) 49–55.
[59] F. Arteaga, Control estadístico multivariante de procesos con datos faltantes mediante análisis de componentes principales, Ph.D. thesis, Universidad Politécnica de Valencia (2003).
[60] N. D. Tracy, J. C. Young, R. L. Mason, Multivariate control charts for individual observations, Journal of Quality Technology 24 (2) (1992) 88–95.
[61] J. Camacho, Missing-data theory in the context of exploratory data analysis, Chemometrics and Intelligent Laboratory Systems 103 (1) (2010) 8 – 18.
[62] F. Arteaga, A. Ferrer, Dealing with missing data in MSPC: several methods, different interpretations, some examples, Journal of Chemometrics 16 (8-10) (2002) 408–418.
[63] S. Wold, Cross-validatory estimation of the number of components in factor and principal components models, Technometrics 20 (4) (1978) 397–405.
[64] S. Becker, A. Feist, M. Mo, G. Hannum, B. Palsson, M. Herrgard, Quantitative prediction of cellular metabolism with constraint-based models: The COBRA toolbox, Nature Protocols 2 (3) (2007) 727–738.
[65] J. Camacho[link].
URL http://wdb.ugr.es/∼josecamacho/