UNIVERSITA' DEGLI STUDI DI PADOVA

DIPARTIMENTO DI INGEGNERIA INDUSTRIALE

CORSO DI LAUREA MAGISTRALE IN CHEMICAL AND PROCESS ENGENEERING

Tesi di Laurea Magistrale in

Chemical and Process Engineering

Enhancing the understanding of CHO cell culture metabolic traits through integrated first-principle modelling and data-based parameter estimation

Relatore: Prof. Pierantonio Facco

Correlatori: Dr. Gianmarco Barberi

Dr. Paloma Diaz-Fernandez

Laureando: Tamiazzo Edoardo

Anno Accademico 2022/2023

Foreword

The fulfillment of the research results included in this Thesis involved the intellectual and financial support of many people and institutions, to whom the author is very grateful.
Most of the research activity that led to the results reported in this Dissertation has been carried out at CAPE-Lab, Computer-Aided Process Engineering Laboratory, at the Department of Industrial Engineering of the University of Padova (Italy), under the supervision of Prof. Pierantonio Facco, and with the great help Gianmarco Barberi. Part of the work resulted from a collaboration with Mrs. Paloma Diaz-Fernandez from GlaxoSmithKline Process Engineering & Analytics and Biopharm Process Research –Stevenage (UK).
Financial support has been provided by the University of Padova

Abstract

Monoclonal antibodies are biopharmaceuticals used to treat diseases such as cancer, leukemia, and asthma. The best host platform for growing these antibodies is mammalian Chinese hamster ovary (CHO) cells; however, the complex and variable nature of this system poses a significant challenge to product and process development.

This Thesis proposes a new methodology to understand how cell metabolism is related to chemical-physical and biological phenomena occurring into the cell culture to understand the connections between those phenomena and cells metabolism. To this purpose, an improved CHO cell first-principle model is proposed. This model is based on several Literature models (Botton et al., 2022; Jimenez del Val et al., 2016; Kontoravdi et al., 2010b) and main improvements are related to the introduction of Glutamate contribution and refinement of the Lactate consumption term. The parameters of this model are estimated using process data (e.g., Viable cell concentration, Product Titer, etc.) that describe the macroscopic behavior of the system and allow to fit a descriptive model of the culture. Metabolomics data are then integrated into this framework to provide new insights and reveal the connections between the phenomena occurring into the culture and cellular metabolism. Metabolomics data provide a microscopic perspective of what happens into the cells and are essential for investigating biological information at the metabolite level. A multivariate latent-variable regression techniques, namely partial least squares (PLS), is used to relate the dynamic behavior of cell metabolism and the most significant chemical-physical and biological phenomena, by estimating the most important model parameters, identified through a in-depth sensitivity analysis, from metabolomics dynamics data. Accordingly, it is possible to determine which metabolites is related to a specific chemical-physical and biological phenomenon.

The proposed framework is applied to the industrial case study of CHO cell culture for the production of monoclonal antibodies at the scale of microreactors from GlaxoSmithKline, Process Engineering & Analytics and Biopharm Process Research (Stevenage, U.K.).

The results show that cell metabolism is strongly related to the model parameters, with an average determination coefficient of 72% in validation. Furthermore, the outcomes allow to understand that cell growth is mainly related to the metabolism of Arginine and Taurine, while antibody production is mainly related to higher Thiamine monophosphate content and Glutamate metabolism.

Table of Contents

Introduction

Monoclonal antibodies (mAbs) are large, lab-grown proteins that mimic the innate disease-fighting capabilities of the immune system. These molecules are engineered to combat invasive threats, including viruses and cancer cells and are precisely designed to target and neutralize pathogens and foreign molecules by binding to specific antigens (Castelli et al., 2019). This fundamental role makes mAbs crucial for the treatment of a wide range of diseases, including cancer, leukemia, asthma, macular degeneration, arthritis, Crohn's disease and post-transplant complications (Quinteros et al., 2017). As a result, mAbs are among the top-selling biologicals and occupy an important position in the biopharmaceutical market (Lu et al., 2020).

Monoclonal antibodies are commonly produced using mammalian cells in a medium that contains the necessary macro and micronutrients. Among these cells, Chinese Hamster Ovary (CHO) culture is the preferred platform for fed-batch mAbs production (Walsh & Walsh, 2022). The development of a process for the effective production of mAbs requires significant resources and time. Therefore, optimizing research and production processes can significantly accelerate the production of high-quality products (Barberi, 2023). However, open issues remain in the efficient development of mAbs, particularly related with the complexity of the cell culture system, where the chemical-physical and biological phenomena occurring into the culture are typically poorly understood. To address these challenges, the biopharmaceutical industry and the scientific community have focused on harnessing the capabilities of Industry 4.0 (Barberi, 2023). This paradigm offers novel approaches to utilize the significant quantities of physical, chemical, and biological data collected during the process. High-throughput systems have become indispensable for gathering extensive data sets, creating the basis for advanced bioprocess modeling (Ahn & Antoniewicz, 2012). In CHO cell culture development, two main categories of data are usually collected: process data, which give insight into the overall behavior of the cell culture through measurements of key parameters and chemical properties (Botton et al., 2022; Facco et al., 2020); and biological (i.e., -omics) data, which offer a microscopic perspective into the internal characteristics and behaviors of the cultured living organisms (Barberi et al., 2022). Among these, metabolomics stands out as crucial factors, concentrating on examining biological information at the metabolite level. Metabolites are vital in different chemical, physical, and biological processes, therefore have significant potential to explain these biological systems. However, working with metabolomic data presents challenges due to the vast number of measurements (ions) and their intricate complexity. Overcoming these issues and harnessing the potential of metabolomic data allows to optimize monoclonal antibody production and provide valuable insights.

The current state of the art in this field characterizes CHO cell cultures with two different approaches. One is the use of first-principles nonlinear models to accurately describe the behavior of such systems (Barberi, 2023; Botton et al., 2022; Jimenez del Val et al., 2016; Kontoravdi et al., 2010). Within these models, the complexity arises from many parameters, each of which is intrinsically linked to chemical, physical, or biological culture phenomena that give them significant physical meaning. Despite their complexity, these models provide the most appropriate approaches for describing the dynamic behavior of mammalian cell cultures in terms of cell growth and metabolic activities (Kyriakopoulos et al., 2018). However, a limitation of this methodology is its inability to explicitly incorporate the amount of metabolic information; additionally, the models are often too simplified to describe accurately the complex culture phenomena.

Other state-of-the-art approaches study the metabolomics information by means of machine learning techniques (Barberi, 2023; Barberi et al., 2022). In this case, the use of multivariate techniques and neural networks allows to build models in which the behavior of CHO cells is described using the metabolites and other omics data. This type of model proves to be very efficient. However, in the existing Literature, there are no explicit efforts to correlate metabolism with the culture chemical, physical, and biological phenomena occurring into the biopharmaceutical process. In particular, correlations between cellular metabolism and cell phenomena remain unexplored due to a lack of available techniques for such investigations. In this context, the main objective of this work is to introduce an innovative way to integrate cell metabolism information from metabolomics data into the framework of first-principles models for CHO cell lines. This allows to simultaneously leverage the descriptiveness of first-principles models, where model parameters represent cell phenomena, with machine learning. In particular, this integration is carried out by building a new accurate first-principles model in which the role of each parameter is strongly associated to a chemical-physical and biological phenomenon. After that the correlation between the most important model parameters (identified through sensitivity analysis) and metabolomics dynamics is explored through multivariate latent-variable regression techniques to gain valuable insights into processes and uncover the intricate connections between the phenomena occurring within the biological system and cell metabolism.

Chapter 1

Materials and methods

In this Chapter, the mathematical methodologies used in this Thesis are described. These techniques, for both first-principle and data based modelling, include the methods for first-principle model structural identifiability and the sensitivity analysis of models' parameters and methodologies of unsupervised and supervised machine learning, such as multivariate statistical models like Principal Component Analysis (PCA) and Partial Least-Squares (PLS).

1.1 First-principle models

First-principles models are nonlinear systems of equations that provide a comprehensive understanding of the dynamic behavior of cellular phenomena. In these models, many parameters are used to define the behavior of the cell lines and these parameters are estimated from process data. The structure of the model must be verified by assessing the structural identifiability of the model parameters. In addition, the most important cell phenomena can be selected by ranking the most characteristic parameters and performing a sensitivity analysis.

1.1.1 Structural Identifiability

Structural Identifiability is an essential prerequisite for parameter estimation (Godfrey, 1999). A model is said to be structurally identifiable if it is possible to determine the values of its parameters from measurements of the model outputs. The method for Structural Identifiability used in this Thesis is called STRIKE-GOLDD (STRuctural Identifiability taKen as Extended-Generalized Observability with Lie Derivatives and Decomposition) (Villaverde et al., 2016).

1.1.2 Procedure for model Structural Identifiability

To understand the procedure behind the structural identifiability of model parameters, consider a generic dynamic model M.

$$M: \begin{cases} \dot{\boldsymbol{x}} = \boldsymbol{f}[\boldsymbol{x}, \boldsymbol{h}, \boldsymbol{d}] \\ \boldsymbol{y} = \boldsymbol{g}[\boldsymbol{x}, \boldsymbol{d}] \\ \boldsymbol{x_0} = \boldsymbol{x}(t_0, \boldsymbol{d}) \end{cases}, \qquad (1.1)$$

where **f** and **g** are vector functions, **d** is the vector of s parameters and **h** is the vector of inputs, **y** is the vector of the measurable outputs and $\mathbf{x_0}$ is the vector of initial conditions. Mathematically, a parameter d_i is structurally identifiable if for any value of d_i the following property holds:

$$y(t, \widehat{d}_i) = y(t, d_i^*) \rightarrow \widehat{d}_i = d, \tag{1.2}$$

Namely, parameter d_i is structurally identifiable if it can be uniquely determined from the system outputs. Consequently, a model is said to be structurally identifiable if all parameters are structurally identifiable.

Within this framework, parameter identifiability can be considered as an augmented observability property (August & Papachristodoulou, 2009). A system is (locally) observable at a state x_A if there exists a neighborhood of x_A such that every other state in the neighborhood is distinguishable from x_A. Two states $\mathbf{x}_A \neq \mathbf{x}_B$ are said to be distinguishable when there exists some input **h** such that $\mathbf{y}(t, \mathbf{x}_A, \mathbf{h}) \neq \mathbf{y}(t, \mathbf{x}_B, \mathbf{h})$. For a nonlinear system it is possible to obtain information about the states **x** from its outputs **y** by calculating the Lie derivatives of the output function **g**. The Lie derivative of **g** with respect to **f** is defined by:

$$L_f \boldsymbol{g} = \frac{\partial \boldsymbol{g}}{\partial \boldsymbol{x}} \boldsymbol{f}(\boldsymbol{x}, \boldsymbol{u}), \tag{1.3}$$

For a system with δ-states and m-outputs $\partial\mathbf{g}/\partial\mathbf{x}$ is an $[m \times \delta]$ matrix and $L_f \boldsymbol{g}$ is an $m \times 1$ vector. Additionally, the i^{th} order Lie derivative are recursively defined as:

$$L_f^i \boldsymbol{g} = \frac{\partial L_f^{i-1} \boldsymbol{g}}{\partial \boldsymbol{x}} \boldsymbol{f}(\boldsymbol{x}, \boldsymbol{u}), \tag{1.4}$$

The nonlinear observability matrix $\mathbf{O_I}$ can then be built as follow:

$$\boldsymbol{O_I} = \begin{pmatrix} \frac{\partial}{\partial \boldsymbol{x}} \boldsymbol{g} \\ \frac{\partial}{\partial \boldsymbol{x}} (L_f \boldsymbol{g}) \\ \vdots \\ \frac{\partial}{\partial \boldsymbol{x}} (L_f^{n-1} \boldsymbol{g}) \end{pmatrix}, \tag{1.5}$$

Based on that, the observability rank condition states that a system is locally observable around $\mathbf{x_A}$ if $rank(\mathbf{O_I}(\mathbf{x_A})) = \delta$.

The described identifiability problem can be transferred to the framework of observability if the s parameters **d** are considered as additional states with no-dynamics $(\dot{\mathbf{d}} = 0)$. In this way, the state variable vector can be augmented by including model parameters, $\tilde{\mathbf{x}} = [\mathbf{x}, \mathbf{d}]$, and the obtain matrix $\mathbf{O_I}(\tilde{\mathbf{x}})$ is referred as generalized observability-identifiability matrix. The generalized Observability-Identifiability Condition (OIC) can be defined so, if the system satisfies $rank(\mathbf{O_I}(\mathbf{x_A})) = \delta + s$, then, it is (locally) observable and identifiable.

1.1.3 Structural Identifiability limitations and solutions

In practice, verifying model structural identifiability by checking the OIC is often computationally inefficient (or even infeasible). Especially for large models, constructing $\mathbf{O_I}$ and computing its rank is a very demanding task. For this reason, *i) limiting the number of calculated Lie derivatives* and *ii)* decomposing the original system into several sub models, are two techniques that can be applied to reduce the computational cost and access large model structural identifiability.

1. The main limitation for large and complex model is the calculation of high order Lie derivatives the rank of the resulting $\mathbf{O_I}$ matrix. The $\mathbf{O_I}$ matrix is built by vertically stacking all the $\delta + s$ Lie derivatives sub-matrices. resulting in final dimension of full matrix $\mathbf{O}_I$ of $\big(m \cdot (\delta + s)\big) \times (\delta + s)$. However, it may not be always necessary to calculate all the Lie derivatives in order to test whether $\mathbf{O_I}$ is full rank, since this may be achieved with a lower number of derivatives. The minimum number of Lie derivatives for which the matrix may be full rank is $n_d = [(\delta + s)/m - 1]$ (Villaverde et al., 2016). For this reason, to reduce the computational cost it is possible to build $\mathbf{O_I}$ by calculating Lie derivatives until the number of rows is greater or equal to n_d. Once n_d is reached, the addition of a new Lie derivative is followed by the calculation of the rank. This process is repeated until the maximum number $\delta + s - 1$ is reached, or until adding a new Lie derivative does not increase the matrix rank; in both cases no further derivatives are necessary. At that point, if $\mathbf{O_I}$ is fullrank the corresponding model is observable and identifiable;
2. another solution is to decompose the model M into sub-models $\{M_1, M_2, \dots\}$ that require few Lie derivatives for their analysis. Studying the structural identifiability of each sub-model is equivalent of study the structural identifiability of the entire model (Villaverde et al., 2016). Each sub-model M_{sub} includes a subset of the model states, $\mathbf{x_{sub}}$. Its outputs, $\mathbf{y_{sub}}$, are the outputs of M which are functions of at least one state included in $\mathbf{x_{sub}}$. The submodel parameters and inputs are those appearing in the equations of $\mathbf{x_{sub}}$ and $\mathbf{y_{sub}}$.
   This sub models can be found by optimization. For each sub-model M_i a subset of states in M is selected by performing an optimization where n_d is minimized.

$$\min_{ex} n_d(ex), \tag{1.6}$$

   where $\mathbf{ex} = \{ex_1, ex_2, \dots, ex_n\}$ is a binary vector of size n, whose entries $ex_j \in \{0, 1\}$ denote inclusion ($ex_j = 1$) or exclusion ($ex_j = 0$) of the corresponding state. The combinatorial optimization is performed with the Variable Neighborhood Search metaheuristic (Egea et al., 2014).

In both cases if $\mathbf{O_I}$ is not full rank, the model is not identifiable and the OIC does not inform about which parameters are identifiable and which are not. However, if deleting the i^{th} column of the $\mathbf{O_I}$ does not change its rank, then the corresponding i^{th} state (parameter) is non-

identifiable). This fact can be exploited to determine which of the parameters in an unidentifiable model are identifiable. After the matrix rank has been calculated, each of the columns in $\mathbf{O_I}$ is removed one by one and the rank is recalculated. In this way even if the system Structural Identifiability is not achieved the identifiability of each of the parameters is evaluated

1.1.4 Sensitivity Analysis

Sensitivity analysis is defined as the study of how uncertainty in the output of a model can be attributed to different sources of uncertainty in the model input. Performing model's parameters sensitivity analysis can by a decision driven by different aspects:

- factor prioritization: prioritize factors that are most deserving of further analysis or measurement;
- factor ranking: order the factors from the most important to the least important;
- model simplification: Fixing or simplifying some factors or compartments of the model.

In this Thesis two sensitivity approaches are used: *i)* the Elementary Effect Test *ii)* the Variance Based Sensitivity Analysis. Particularly, the sensitivity analysis is performed to identify a set of parameters that most accurately capture the overall behavior of the cell culture.

1.1.5 Elementary Effect Test

The Elementary Effect Test (EET) is a method to find an approximate sensitivity information (Saltelli et al., 2008a). The Elementary Effect index (EE_i) is an average of derivates performed at different points sampled over the space of factors. Considering s parameters the EE_i of the i^{th}-input parameter is defined as:

$$EE_i = [y(d_1, d_2, \dots, d_i + \Delta, \dots, d_s) - y(d_1, \dots, d_s)] \,/\, \Delta, \qquad (1.7)$$

where d_i is the i-th parameter, y is the response variable and Δ is the parameter variation.
To effectively calculate the value of EE_i, the sampling space (Ω) is divided in a grid and the absolute value of EE_i can be computed at different grid points.

1.1.6 Sensitivity indexes for the Elementary Effect Test

Results of the Elementary Effect analysis are collected in the matrix of the Elementary Effects (**EE**). From that two indices one can calculate:

- μ_i, mean of the EE_i. These describes the mean effect in the response by changing the parameter across Ω;
- σ_i, standard deviation of the EE_i. It calculates how much the value of μ_i, vary in the range of the variation of the single parameter.

The value of μ_i, assesses the overall influence of the parameter on the output obtained by averaging the value of the EE_i. In this work, the use of the mean μ_i is replaced by μ_i^*, which is

defined as the mean of the absolute values of the Elementary Effects. The use of μ_i^* is convenient as it solves the problem of the Type II error (i.e., failing the identification of a parameter of considerable influence on the model) to which the μ-value can be exposed (Campolongo et al., 2007). Type II errors might occur when the distribution contains both positive and negative elements. In this case, computing parameters' mean determine that some effects may cancel each other out, thus producing a low mean value even for an important factor. The value of μ_i^* is a very useful index (Saltelli et al., 2008a) as:

- provide a semi-quantitively measure to rank factors;
- is numerically efficient;
- is a good approximation for more complex sensitivity index.

The second calculated index is the standard deviation σ_i, which estimates the ensemble of higher order effects of the parameter. If for a parameter d_i a high value of σ_i is obtained compared to d_i, then the respective EE_i differ significantly from each other. This means that the values of EE_i are strongly influenced by the choice of the sample points at which they are computed and thus by the choice of the values of the other parameters.

1.1.7 Latin Hypercube Sampling

The value of EE_i is obtained by sampling the parameters values in an iterative procedure. This operation can be performed by the introducing the Latin Hypercube Sampling (LHS) that reduces the computational cost of the analysis (McKay et al., 1979). This sampling methodology is a sort of stratified Monte Carlo sampling procedure. Practically, in a case of a one-dimensional LHS to generate a random sample with Γ data points, *i)* the parameter cumulative density function is divided into Γ equal intervals at same probability. Then *ii)* in each interval a point is selected randomly, and this give Γ different points. This rationale can be extended to two independent parameters d_1 and d_2 (two-dimensional LHS). It is possible to generate two one dimensional sample for d_1 and d_2 separately. Once we have two lists of samples, they are combined into two-dimensional pairs. The same procedure can be extended to a larger number of parameters. Compared to a classical Montecarlo approach, LHS has several advantages:

- tends to be spread more uniformly the samplings along Ω (Morris, 1991);
- is more efficient and less time consuming than a Monte Carlo simulation (Atangana, 2017).

1.1.8 Variance Based Sensitivity Analysis

Variance Based Sensitivity Analysis (VBSA) is a global sensitivity analysis to assess the sensitivity of the response to parameter variation (Saltelli et al., 2008b). Consider a model in the form of $Y = \mathbf{f}(d_1, d_2, \dots, d_k)$ in which each d_i is a parameter that has a non-null range of variation. The concept behind VBSA is in the investigation of Y total variance (i.e., unconditional variance) indicated with $V(Y)$ when some parameters are kept fixed. If a

parameter is frozen to a specific value $d_i = d_i^*$, then the resulting variance $V(Y|d_i = d_i^*)$ will be certainly smaller compared to $V(Y)$. Hence, the smaller $V(Y|d_i = d_i^*)$ is, the greater is the influence of d_i on the variance of Y. However, $V(Y|d_i = d_i^*)$ is affected by the choice of d_i^*, so this concept can be generalized by averaging the measure of $V(Y|d_i = d)$ with different d_i^* points, so that the dependence on specific parameter value will disappear. The average of the measurements of $V(Y|d_i = d_i^*)$ at different d_i^*values is referred as $E_{d_i}\left(V_{d_i}(Y|d_i)\right)$ and this is always lower or equal to $V(Y)$.

1.1.9 First order Variance Based Sensitivity Index

The conditional variance $V_{d_i}(E_{d_i}(Y|d_i))$ is defined as:

$$V_{d_i}\left(E_{d_i}(Y|d_i)\right) = V(Y) - E_{d_i}(V_{d_i}(Y|d_i), \quad (1.8)$$

Lower value of $E_{d_i}(V_{d_i}(Y|d_i)$ means that d_i is an important parameter, consequently, the higher $V_{d_i}\left(E_{d_i}(Y|d_i)\right)$ is, the higher is the importance of d_i.

The conditional variance is called first order effect of d_i on Y and the associated sensitive index is:

$$S_i = \frac{V_{d_i}\left(E_{d_i}(Y|d_i)\right)}{V(Y)} \in [0;1], \quad (1.9)$$

High value for S_i means that d_i is an important parameter, however no conclusive statement can be done if S_i is very low. In fact, the first order index considers only the main effect of a factor, but nothing says on its total effect considering interactions with other parameters.

1.1.10High order Variance Based Sensitivity Index

The first ordered sensitivity index is obtained by conditioning one parameter at time, however this concept can be generalized to any higher order. Let's consider for simplicity to use two conditioned factors instead of one, the conditioned variance is calculated as:

$$V\left(E\left(Y\middle|d_i, d_j\right)\right) = V_i + V_j + V_{ij}, \quad (1.10)$$

here V_i is the conditioned variance for the i-th parameter, V_j is the conditioned variance for the j-th parameter and V_{ij} considered parameters interaction effect.

In Equation (*1.10*), V_i and V_j can be computed from Equation 1.8 and $V(E(Y|d_i, d_j))$ is calculated as the conditioned variance by keeping fixed two parameters:

$$V(E(Y|d_i, d_j)) = V(Y) - E_{d_i}(V_{d_i}(Y|d_i) - E_j(V_{d_j}(Y|d_j), \quad (1.11)$$

then:

$$V_{ij} = V(E(Y|d_i, d_j)) - (d_i + d_j), \quad (1.12)$$

V_{ij} is the interaction term between d_i and d_j and capture the part of response that cannot be written as superposition of the two-separate effect of d_i and d_j.

If the rationale of conditioned variances is expanded to a general order, all the factors except for one can be conditioned. In that case, the conditioned variance $V(E(Y|\mathbf{d} - \{d_i\}))$ captures all terms of any order that do not include parameter d_i and the associate sensitivity index is equal to $V(E(Y|\mathbf{d} - \{d_i\}))/\, V(Y)$. Additionally, the sum of all possible sensitivity terms must be equal to one, as:

$$\sum_i S_i + \sum_i \sum_{j>i} S_{ij} + \sum_i \sum_{j>i} \sum_{l>j} S_{ijl} + \cdots + S_{123\ldots k} = 1\,, \quad (1.13)$$

where S_i is the first order sensitivity index, S_{ij} is the second order sensitivity index, $S_{ijl}$is the third order sensitivity index and $S_{123\ldots k}$ is the k-order sensitivity index.

Then the difference $1 - V(E(Y|\mathbf{d} - \{d_i\}))/\, V(Y)$ must be made up of all terms of any order that include d_i, namely the total effect of d_i. Based on that we define the total-sensitivity index of d_i (i.e., Sobol's index) as:

$$S_{T_i} = 1 - \frac{V\left(E\left(Y|d - \{d_i\}\right)\right)}{V(Y)}\,. \quad (1.14)$$

Calculation of the sensitivity indexes are performed by iteratively calculating the conditional variances (value of $V(Y|d_i = d_i^*)$) for different parameter values. The values of the parameters are sampled by LHS. The calculation of the indexes is performed with the SAFE MATLAB Toolbox (Pianosi et al., 2015) in which main and total effects are calculated as explained above (Homma & Saltelli, 1996; Saltelli et al., 2008a, 2008b, 2010).

1.1.11Limitations of Variance Based Sensitivity Analysis

A good characterization of sensitivity of a system is given by the total set of first-order terms plus the total effects. Additionally, if the total sensitivity index of a parameter, $S_{Ti} = 0$ then this is a necessary and sufficient condition for d_i to be non-influent in the model (this does not hold simply for the first order sensitivity index S_i).

VBSA is more efficient with respect to the EET. However, it has several problems:

- computationally heavy. VBSA lead to higher computational burden. To decrease the time required for the simulation EET can be used. This prunes the number of factors as EE_i significance is a necessary condition for the parameter to be also significant for the VBSA;
- if the output distribution is multi-modal or if it is highly skewed, using variance as a approximation of uncertainty may lead to contradictory result (Pianosi & Wagener, 2015). There are different reported cases (Borgonovo et al., 2011; Liu et al., 2006) in which highly

skewed distributions have proved to not be efficient in the application of VBSA. If the curve is highly skewed, results need to be interpreted comparing them with the ones obtained from the EET;

- sensitivity index is lower than zero. Any order sensitivity is limited between 0 and 1, where1 indicates that parameter is the only influential one. However, in this work, the used method calculate the Sobol's indexes (Saltelli et al., 2010) does not guarantee that $V\left(E(Y|d_i)\right)$ is always positive. Accordingly, sometimes S_i might exhibit negative values. To overcome this numerical problem, is necessary to increase the numbers of simulation to bring this value as close to zero as possible and if any negative value occurs, is possible to keep it as zero at the condition that its confidence interval contains the zero.

1.2 Machine learning

In this Thesis several machine learning techniques are used to perform several tasks. K-means analysis is used to cluster the different cell lines and identify common behaviors. Multivariate methods are statistical models used for dimensionality reduction, data interpretation and visualization, correlation analysis, and regression/classification. Two multivariate methods are implemented: *i)* Multiway Principal Component Analysis and *ii)* Multiway Partial Least Squares. In both cases, the dataset analyzed by multivariate models are preprocessed: dataset are mean-centered (i.e., by removing the column-wise mean value), auto scaled (i.e., by removing the column-wise mean value and scaling for the column-wise standard deviation) or pareto-scaled (i.e., by removing the column-wise mean value and scaling for the square root of column-wise standard deviation).

Particularly in the Partial Least Squares, the importance of each regressor variable (i.e., in **X**) for the prediction of **Y** is quantified through several indices. Three different indices are used to assess the importance of predictor variables: *i)* the regressor coefficients, *ii)* VIP scores and *iii)* the Selectivity Ratio. These indexes are utilized to select the most important ions for building the multivariate models. This ion discrimination is achieved via a robust and computationally intensive backward iterative elimination method (Barberi, 2023) Furthermore, the VIP value and regressor coefficients are also applied to differentiate the most significant metabolites in each PLS model.

1.2.1 k-means clustering

The k-means method is a widely used clustering technique that aims at minimizing the average squared distance between data points within the same cluster. While it offers no explicit guarantees of accuracy in finding the best clusters, its practical appeal lies in its simplicity and efficiency. Mathematically, the algorithm follows a straightforward process in four steps:

1. initialization: select k initial cluster centroids $c = \{c_1, c_2, \dots, c_k\}$;

2. assignment: for each $l \in \{1, \dots, k\}$ set the cluster C_l to be the set of points closer to c_l accorindg to a selected distance metric (i.e., measurement of dissimilarity between observation) which, for example, can be the euclidean distance;
3. update: for each $l \in \{1, \dots, k\}$ recompute the centroids so that c_l is set to be the center of mass of all points in C_l;
4. iterate: repeat steps 2 and 3 until convergence is reached, or a specified number of iterations is reached.

1.2.2 Principal component analysis

Principal Component Analysis (PCA; Wise & Gallagher, 1996) is a powerful method for reducing the dimensionality of multivariate data and extracting the main driving forces. PCA decomposes a scaled dataset $\mathbf{X}$ $[N \times V_R]$, consisting of N observations and V_R variables, into A orthogonal principal components (PCs). These principal components indicate the most significant directions of variability within $\mathbf{X}$ and effectively capture the relationships among the V_R variables. In PCA the $\mathbf{X}$ dataset is decomposed as:

$$\mathbf{X} = \mathbf{T}\mathbf{P}^{\mathrm{T}} + \mathbf{E}\,, \tag{1.15}$$

where $\mathbf{T}$ $[N \times A]$ is the score matrix, $\mathbf{P}$ $[V_R \times A]$ is the loading matrix, the superscript T indicates the transpose, and $\mathbf{E}$ $[N \times V_R]$ is the residual matrix, which is minimized in the least-squares sense. The scores represent the projection of the samples onto the space defined by the principal components and illustrate the relationship among the N observations. Meanwhile, loadings describe the pattern of correlations among the V_R variables. Common methods to compute model scores and loadings (i.e., calibration) are singular value decomposition or nonlinear iterative partial least squares. (NIPALS; Geladi & Kowalski, 1986).

In this Thesis, the number A of PCs (i.e., the dimension of the reduced space) is selected through cross-validation. In cross-validation (Wold, 1978), the optimal number of PCs is chosen to minimize the reconstruction error, usually measured by the root mean squared error (RMSE) using a bootstrapping or jackknifing technique. The main model diagnostics to assess the performance of the model are the RMSE and the coefficient of determination R^2. The RMSE is defined as:

$$RMSE = \sqrt{\frac{\sum_{n=1}^{N}(x_n - \hat{x}_n)^2}{N}}\,, \tag{1.16}$$

where x_n is the n-th sample, and $\hat{\mathrm{x}}_n$ is the n-th sample reconstructed by the PCA model. The coefficient of determination quantifies the amount of variability of the original data $\mathbf{X}$ captured by the model, and it is defined as:

$$R^2 = 1 - \frac{\sum_{v=1}^{V}\sum_{n=1}^{N}(x_{n,v} - \hat{x}_{n,v})^2}{\sum_{v=1}^{V}\sum_{n=1}^{N}(x_{n,v} - \bar{x}_v)^2}, \quad (1.17)$$

where $x_{n,v}$ is the value of the v-th original variable for the n-th sample, $\hat{x}_{n,v}$ is the value of the v-th original variable for the n-th sample reconstructed by the PCA model, and $\bar{x}_v$ is the average value of the v-th original variable.

After calibrating a PCA model, a new observation $\mathbf{x}_{NEW}$ $[1 \times V_R]$ can be projected onto the PCA model to compare it with the calibrated observations and evaluate if its fit to them. This projection is performed as:

$$\mathbf{t}_{\mathrm{NEW}} = \mathbf{x}_{\mathrm{NEW}}\mathbf{P}, \quad (1.18)$$

where $\mathbf{t}_{NEW}$ $[1 \times A]$ represents the score vector for the new observation. To assess how well the model describes an observation, to detect potential outliers, and to determine the impact of an observation on the overall model, sample diagnostics such as Hotelling's T^2 and the squared prediction error (SPE) can be computed. Hotelling's T^2 quantifies the distance between the projection of an observation and the origin of the reduced space, typically indicating the magnitude of the deviation of a particular sample from the average conditions of the calibration data set. The Hotelling's T^2 for a given observation n is defined as:

$$\mathrm{T}_{\mathrm{n}}^2 = \mathbf{t}_{\mathrm{n}}\boldsymbol{\Lambda}^{-1}\mathbf{t}_{\mathrm{n}}^{\mathrm{T}}, \quad (1.19)$$

where $\mathbf{t}_n$ is the score vector of the n-th observation, and $\boldsymbol{\Lambda}^{-1}$ $[A \times A]$ is a diagonal matrix collecting the inverse eigenvalues. The SPE quantifies the mismatch between an observation and its reconstruction through the PCA model. Large values of SPE identify observations with a correlation structure different than the one in the calibration dataset. The SPE for a given observation ni is defined as:

$$\mathrm{SPE}_{\mathrm{n}} = \mathbf{e}_{\mathrm{n}}\mathbf{e}_{\mathrm{n}}^{\mathrm{T}}, \quad (1.20)$$

where $\mathbf{e}_n = \mathbf{x}_n - \hat{\mathbf{x}}_n$ is the residual vector of the w-th observation. It is possible to construct confidence limits for both Hotelling's T^2 and SPE (Nomikos & MacGregor, 1995a) to detect potential outliers. These statistical calculations assume that the data used to construct the model are independent and normally distributed. This assumption results in scores that follow a multi-normal distribution and residuals that resemble white noise. The confidence limit on the Hotelling's T^2, T_{lim}^2, is calculated as:

$$T_{lim}^2 = \frac{A(N-1)}{(N-A)} F_{A,N-A,\alpha}, \quad (1.21)$$

where $F_{A,N-A,\alpha}$ is the critical value of a F-distribution with A and $N - A$ degrees of freedom at the significance level α.The confidence limit on the SPE, SPE_{lim}, is calculated as:

$$SPE_{lim} = \frac{\sigma_{SPE}^2}{2\mu_{SPE}} \chi^2_{2\mu_{SPE}^2/\sigma_{SPE,\alpha}^2}, \quad (1.22)$$

where $\chi^2_{2\mu_{SPE}^2/\sigma_{SPE,\alpha}^2}$ is the critical value of a χ^2-distribution with $2\mu_{SPE}^2/\sigma_{SPE}^2$ degrees of freedom at the significance level α, μ_{SPE} is the average, and σ_{SPE} is the variance of the SPE distribution.

1.2.3 Partial Least-Squares

Partial Least-Squares Regression (PLS; Wise & Gallagher, 1996) is a linear multivariate regression method used to capture the joint correlations between a matrix of predictors and a matrix of responses, and to predict new responses using a set of new predictors. PLS identifies the maximum direction of covariance between a scaled matrix of predictors $\mathbf{X}$ $[N \times V]$ and a scaled matrix of responses $\mathbf{Y}$ $[N \times R]$ containing R responses. This technique projects both $\mathbf{X}$ and $\mathbf{Y}$ into a reduced space defined by A latent variables (LVs) using the following approach:

$$\mathbf{X} = \mathbf{T}\mathbf{P}^{\mathrm{T}} + \mathbf{E}, \quad (1.23)$$

$$\mathbf{Y} = \mathbf{T}\mathbf{Q}^{\mathrm{T}} + \mathbf{F}, \quad (1.24)$$

$$\mathbf{T} = \mathbf{X}\mathbf{W}(\mathbf{P}^{\mathrm{T}}\mathbf{W})^{-\mathbf{1}}, \quad (1.25)$$

where $\mathbf{P}$ $[A \times V_R]$ and $\mathbf{Q}$ $[A \times R]$ are the loading matrices, $\mathbf{T}$ $[N \times A]$ is the score matrix, $\mathbf{E}$ $[N \times V_R]$ and $\mathbf{F}$ $[N \times R]$ are the residual matrices of $\mathbf{X}$ and $\mathbf{Y}$, respectively (minimized in a least-square sense), and $\mathbf{W}$ $[N \times A]$ is the weight matrix used for the calculation of the scores. The inclusion of weights is crucial to maintain orthogonality within the latent variable (LV) scores and to identify the direction of maximum correlation within the standardized versions of $\mathbf{X}$. It's important to note that the scores $\mathbf{T}$ and loadings $\mathbf{P}$ in a PLS model are different from those in a PCA model. Calculating model scores, loadings, and weights (referred to as calibration) typically involves iterative techniques such as NIPALS algorithm (Wold et al., 2001). PLS can be used for predicting a response variable $\hat{\boldsymbol{y}}$ $[1 \times R]$ from a set of new predictors $\mathbf{x}_{NEW}$ $[1 \times V_R]$ according to:

$$\hat{\mathbf{y}} = \mathbf{x}_{\mathrm{NEW}}\mathbf{W}(\mathbf{P}^{\mathrm{T}}\mathbf{W})^{-\mathbf{1}}\mathbf{Q}^{\mathrm{T}}, \quad (1.26)$$

The number of LVs can be selected similarly to PCA (Section 1.2.2) with the cross-validation.

1.2.4 Multiway modeling by batch-wise unfolding

Multiway multivariate modeling (Nomikos & MacGregor, 1994) is a technique used to deal with complex data sets organized as multidimensional matrices, where one of the dimensions is often related to time (indicating temporal variability in the data). Multiway modeling consists in unfolding the multidimensional data $\underline{\mathbf{X}}$ $[N \times V_R \times T]$ (where T is the number of time instants in which V variables are collected for N batches) followed by the decomposition with a standard multivariate model. In this Thesis the data unfolding procedure followed the rules of the batch-wise unfolding (Nomikos & MacGregor, 1995b) and are schematically shown in Figure 1.1 Data are collected at different time instants (e.g., $\mathbf{X}_t$ $[N \times V_R]$ with $t = 1,2,\dots,T$) and are horizontally concatenated to generate a matrix $\mathbf{X}_{bwu}$ $[N \times V_R \cdot T] = [\mathbf{X}_1\ \mathbf{X}_2\ \cdots\ \mathbf{X}_T]$, which is the batch-wise unfolded version of $\underline{\mathbf{X}}$. In multiway multivariate modeling, loadings play a crucial role in clarifying the correlation among the variables denoted by $\mathbf{X}$ at different time points. This distinction provides insights into the relationships among the dynamics of the variables and their cross-correlations. This Thesis includes the multiway principal component analysis (MPCA) and the multiway of partial least squares (MPLS).

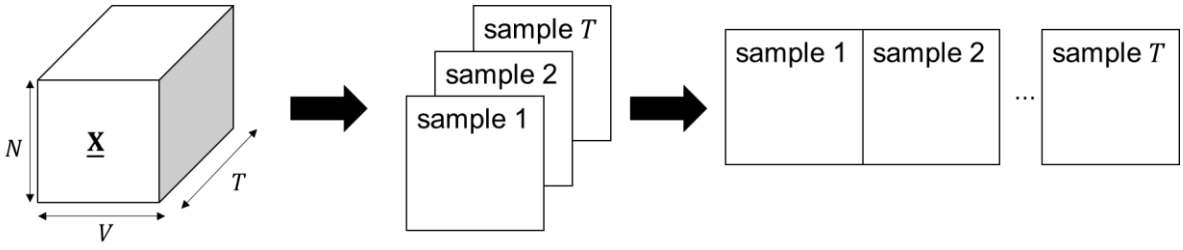


Figure 1.1 *Batch-wise unfolding procedure for multiway multivariate modeling.*

1.2.5 Regression coefficients

In Section 1.2.3 PLS is defined as a linear multivariate regression method that can be used to predict new responses using a new set of predictors. Within this framework, is possible to define: the beta regressors (Wold et al., 2001) as:

$$\mathbf{Y} = \mathbf{XB}\ , \tag{1.27}$$

$$\mathbf{B} = \mathbf{W}(\mathbf{P}^{\mathrm{T}}\mathbf{W})^{-1}\mathbf{Q}^{\mathrm{T}}\ , \tag{1.28}$$

where $\mathbf{B}$ is the matrix of regression coefficient for PLS. Because both the $\mathbf{X}$ and $\mathbf{Y}$ data sets are auto-scaled, higher beta values are indicative of regressors that are more important in influencing the value of the response variable.

1.2.6 VIP

VIP scores summarize the influence of each **X** variable on the PLS model (Chong & Jun, 2005). They are calculated as the weighted sum of squares of the PLS weights, which consider the amount of explained y-variance in each extracted latent variable. For this reason, VIP scores provide a measure useful for selecting which are the variables that contribute the most to the explanation of the y-variance. The VIP score for the v-variable can be calculated as:

$$VIP_v = \sqrt{\frac{V \sum_{v=1}^{A} \left(SS(b_k t_k) \cdot w_{v,a}^2 \right)}{\sum_{k=1}^{h} SS(b_k t_k)}}, \quad (1.29)$$

where $SS(b_k t_k)$ is the variance of the response explained by the a-th LV of the model and $w_{v,a}$ is the weight of the v-regressors and a-the LV. In addition, since the average of the squared VIP scores is equal to 1, the *greater than one rule* can be used as a criterion for variable selection.

1.2.7 Selectivity Ratio

The Selectivity Ratio (SR) (Rajalahti et al., 2009) method is a tool for searching what are the important variables of a multivariate data set in the prediction of a particular property. In particular, SR is defined by the ratio between the explained and the residual (unexplained) variance for each variable in the target projection vector. This target projection utilizes both the predictive ability (regression vector) and the explanatory ability (spectral variance/covariance matrix) for the calculation of the Selectivity Ratio. Given the PLS regression vector:

$$\mathbf{t}_{\mathrm{TP}} = \mathbf{X}\mathbf{b}_{\mathrm{TP}}, \quad (1.30)$$

$$\mathbf{p}_{\mathrm{TP}}^{\mathrm{T}} = \frac{\mathbf{t}_{\mathrm{TP}}^{\mathrm{T}}\mathbf{X}}{(\mathbf{t}_{\mathrm{TP}}^{\mathrm{T}}\mathbf{t}_{\mathrm{TP}})}, \quad (1.31)$$

$$\mathbf{X} = \mathbf{X}_{\mathrm{TP}} + \mathbf{E}_{\mathrm{TP}} = \mathbf{t}_{\mathrm{TP}}\mathbf{p}_{\mathrm{TP}}^{\mathrm{T}} + \mathbf{E}_{\mathrm{TP}}, \quad (1.32)$$

Where $\mathbf{t}_{TP}$ is the vector of target projection scores, $\mathbf{p}_{TP}$ is the vector of target projection loadings and $\mathbf{X}_{TP}$ + $\mathbf{E}_{TP}$ is the target projection model. From that we can calculate explained $v_{expl,i}$ ($\|X_{TP}\|^2$) and residual $v_{res,i}$ ($\|E_{TP}\|^2$) variance for each variable i after the target projection. The ratio of explained to residual variance of a variable is defined as selectivity ratio and represents a measure of a variable's sensitivity:

$$SR_i = \frac{v_{expl,i}}{v_{res,i}}. \quad (1.33)$$

Chapter 2

Case study: monoclonal antibody development

This Chapter gives a general introduction on the biopharmaceutical industry and introduces CHO cell cultures and monoclonal antibodies. The case study presented in this Thesis is then described.

2.1 Biopharmaceutical industry

In medicine, biopharmaceuticals (also known as biologicals or biologics) are drugs and therapeutics derived or synthesized from living organisms. These organisms include microbial, animal, or human cells that have been genetically modified to produce specific biological substances with therapeutic effects. One of the key advantages of biopharmaceuticals over traditional pharmaceuticals is their ability to produce more complex drugs with highly targeted functions (Rader, 2008). This specificity leads to a reduced likelihood of side effects compared to conventional chemically synthesized drugs. The use of living cells as a source for drug production has opened up new possibilities for the treatment and prevention of various diseases. The main biopharmaceutical products are vaccines, cells (such as stem cells), biological tissues, recombinant proteins (such as monoclonal antibodies) and gene therapy drugs.

Monoclonal antibodies (mAbs) dominate the biopharmaceutical market as the best-selling class of biologics (Lu et al., 2020). Within this framework, mammalian cell culture has emerged as the preferred method for the production of recombinant proteins, accounting for 67% of total production (Walsh, 2018). Within this category, Chinese hamster ovary (CHO) cells stand out as the primary cell line responsible for the synthesis of 89% of mAbs produced (Walsh & Walsh, 2022).

2.2 Monoclonal Antibodies

Monoclonal antibodies play a critical role in the treatment of a variety of medical conditions, including breast cancer, leukemia, asthma, macular degeneration, arthritis, Crohn's disease, and transplants (Quinteros et al., 2017)

The primary function of antibodies in living organisms is to eliminate invading pathogens and foreign molecules. They specifically bind to their targets, known as antigens, and form a

complex that is recognized and eliminated by specialized components or cells of the host organism's immune system (Castelli et al., 2019)

Monoclonal antibodies are large Y-shaped proteins composed of two identical chains (one heavy and one light), and two identical light chains linked by disulfide bonds (Chiu et al., 2019). In living organisms, monoclonal antibodies are predominantly produced by secretory B cells, an essential component of the cellular immune system (Gaughan, 2016).

Monoclonal antibodies are typically cultured in mammalian cells in fed-batch reactors. In this system, cells are cultured in a medium containing all the macro- and micronutrients necessary for their growth and survival (Li et al., 2010). The primary carbon sources for the cells are typically Glucose, Glutamate, and Glutamine. In addition, specific amino acids required for the production of the desired product can be supplied via daily boluses.

2.3 CHO cell cultures

Cell culture is the preferred host platform for mAb production. The development of a successful biopharmaceutical molecule encompasses all activities aimed at large-scale production of a biopharmaceutical product. Process development is resource-intensive and time-consuming, and is typically divided into several steps:

- cell generation and engineering. The first stage of process development involves the generation of cell lines that will be responsible for producing the desired monoclonal antibody; during this stage, host cells are genetically engineered to improve or modify various aspects such as product quality and growth characteristics;
- cell line selection and scale-up. Cell line selection involves screening thousands of different cell lines for a limited number of quality attributes (QAs) such as cell growth, specific productivity, and product titer (Facco et al., 2020). This screening process is essential to identify the most promising cell lines that meet the desired critical quality attributes (CQAs). To improve the selection process and optimize cultivation, valuable information from biological profiling, such as metabolomics, can be extracted and used (Barberi, 2023). After the initial screening, only those cell lines that meet the desired CQAs are scaled up, moving from laboratory to production scale. Scale-up is a complex and resource-intensive process; reducing the time to market by early identification of suitable commercial cell lines has a significant impact on the overall economics of the process.
- process characterization. Process characterization is a crucial step in the development of biopharmaceuticals, as it is a requirement for drug approval, specifically as part of the biologic license application;

- media, feed, and process optimization. Another essential step is the optimization of media and feeding schedule to balance cell growth, productivity, and product quality. Equally important is the optimization of the critical process operating parameters (CPPs) to ensure stable and high protein expression while maintaining the desired product quality.

Efficient biomanufacturing and the production of high-quality biopharmaceutical products depend heavily on optimizing these critical steps. However, the advent of Industry 4.0 and the abundance of big data offer new solutions to address these challenges.

The biopharmaceutical industry has recognized the potential opportunities in harnessing the vast amounts of physical, chemical, and biological data generated. For this reason, the implementation of high-throughput systems enables the collection of massive amounts of data and serves as the basis for advanced bioprocess modeling. In CHO cell culture, the main types of available data are:

- process data: the macroscopic behavior of cell cultures is assessed by measuring key process parameters and chemical properties. These measurements are essential for monitoring culture growth and cell metabolism, and for investigating potential factors contributing to a decline in cell health. The most important measurement is the viable cell concentration (VCC), which provides valuable insight into the response of the culture to specific conditions;
- -omics data: biological data provide valuable insights into the internal microscopic properties and behavior of the cultured living organisms. These data revolve around the flow of information within all living organisms, starting from DNA and progressing to mRNA, proteins, and metabolites, ultimately leading to the expression of cell phenotypes (Reel et al., 2021). To capture this informative flow of biological information, -omics data are modelled and categorized based on the source of the information they study. The -omics models provide a comprehensive understanding of cellular processes and help to optimize bioproduction for various applications.

Among the various -omics data, metabolomics focuses specifically on the analysis of biological information at the metabolite level. Metabolomics involves the identification and quantification of all small molecules, called metabolites, involved in the metabolic reactions of a given system. It provides valuable insight into the metabolites which are present in the system and their respective abundances. Mass spectrometry, liquid chromatography-mass spectrometry (LC-MS), and gas chromatography-mass spectrometry (GC-MS) are the typical methods used to perform metabolomic measurements.

2.4 Available data

In this Thesis, an industrial case study concerning the development of mAb at a small bioreactor scale (AMBR15^{TM}) is considered. Data from two runs performed in the AMBR15^{TM} miniature bioreactor system (Sartorius Stedim Biotech, Sartorius AG, Goettingen, Germany) are available. These experiments were performed using GSK proprietary platform process.

The data contains information on $N = 96$ CHO cell lines, all expressing a common product (i.e., monoclonal antibody), grown simultaneously in 48 parallel 15-mL bioreactors for 15 days of culture (i.e., experimental batches).

The production process is carried out in a fed-batch manner. In this system, as the cells grow and consume nutrients, additional nutrients are continuously fed to the bioreactor during the cultivation process (Xu et al., 2023). Glucose and Glutamate are used as the main carbon sources. The process conditions in terms of pH and temperature are the same for all microbioreactors, while the feeding action (daily boluses of nutrients) is different for all cell cultures.

A set of $V_P = 7$ process variables are measured in $T = 7$ time instants during the experimental batch ($t = 1,2, \dots, T$, namely 0, 3, 5, 8, 10, 13 and 15 days). These are:

- Viable Cell Concentration (VCC);
- product titer: concentration of monoclonal antibody in the cell culture;
- Glucose: main nutrient and source of carbon;
- Glutamate: second main nutrient and primary source of nitrogen;
- Glutamine: essential building block generated by Glutamate conversion
- Lactate: product of Glucose anaerobic glycolysis whose accumulation decrease cellular productivity and viability;
- Ammonium: by-product of the cellular activity that cause the death of the cells in the culture;

All the process variables are arranged in a three-dimensional array $\underline{\mathbf{X}}_{\mathbf{P}}\,[N \times V_P \times T] = [96\ cell\ lines \times 7\ process\ variables \times 7\ time\ instants]$, which is defined as process dataset. The unfolded version of this matrix $\mathbf{X_P}\,[N \times TV_P]$ is obtain by batch-wise unfolding $\underline{\mathbf{X}}_{\mathbf{P}}$, following the procedure explained in Chapter 1.

For each CHO cell line, intracellular metabolites are analyzed by using flow injection liquid chromatography-mass spectrometry. The chromatography measurements were performed in negative ionization mode, with a scan range of mass over charge (m/z) from 50 to 1000. The raw data obtained from the analysis are pre-preprocessed through an in-house pipeline, and the identified ions are tentatively assigned as metabolites based solely on accurate mass information. However, due to some limitations, such as isomers or metabolites with masses

falling within the annotation tolerance, some ions are annotated as multiple tentative metabolites.

Metabolomic profiling is performed in $R = 2$ replicates at the same T time points as in the culture analysis. However, metabolomic profiles at time point $t = 2$ are missing, because the number of cells in the cultures was insufficient to perform the analysis. For this reason, the intracellular metabolomic profiles consisting of intensities of $V_I = 4587$, are arranged in four-dimensional arrays $\underline{\mathbf{X}}_{\mathbf{ic}}\,[N \times V_I \times (T-1) \times R] = [96\,cell\ lines \times number\ of\ ions \times (7-1)\,time\ points \times 2\,replicates]$. Initially $\underline{\mathbf{X}}_{\mathbf{ic}}$ is divided between the two replication matrixes: $\underline{\mathbf{X}}_{\mathbf{ic,1}}$ and $\underline{\mathbf{X}}_{\mathbf{ic,2}}[N \times V_I \times (T-1)]$. Then both matrices are unfolded, and the unfolded version of these matrices $\mathbf{X}_{\mathbf{ic,1}}$ and $\mathbf{X}_{\mathbf{ic,2}}$ $[N \times V_I(T-1)]$ are obtained by batch-wise unfolding respectively $\underline{\mathbf{X}}_{\mathbf{ic,1}}$ and $\underline{\mathbf{X}}_{\mathbf{ic,2}}$, following the procedure explained in Chapter 1.

In addition, to ensure reliable analysis, ions with more than 20% missing intensities are removed from the data set. For the remaining missing data, a missing data imputation technique is used (Barberi, 2023; Troyanskaya et al., 2001) to infill the missing measurements. In this method, the missing values are imputed by calculating the weighted average intensity of $K_{miss} = 15$ metabolites that have intensity profiles similar to the metabolite of interest.

Chapter 3

CHO cell cultures modelling: analysis and improvement

In this Chapter, a state-of-the-art model describing CHO cell cultures is analyzed and improved by refining the role of Glutamate and Lactate. The model structure is first assessed by the identification of its parameter. A sensitivity analysis is then performed to retrieve a ranking of the parameters importance to select the subset of the most characterizing parameters whose values are estimated in a step-by-step procedure.

3.1 Framework of the project and research objectives

Monoclonal antibodies serve as essential therapeutic proteins produced by mammalian cell culture, specifically using CHO cells (Walsh & Walsh, 2022). The biopharmaceutical industry and the scientific Literature demonstrate an impressive level of activity in this field, making it a highly trending topic (Walsh & Walsh, 2022). However, despite its importance, CHO cell culture faces numerous challenges and complexities. The system exhibits remarkable biological variability and poorly understood phenomena that contribute to the issues of this critical process.

In our case study, two types of data are available: process data, which describe the overall behavior of the system and are related to the physical phenomena that govern it, and metabolomic information, which provides insight into the metabolic characteristics and internal microscopic properties of the living organisms used in the process. Furthermore, the metabolites within this system are undoubtedly associated with various chemical-physical and biological phenomena. However, working with metabolomic data and drawing meaningful conclusions can be challenging due to the high number of available measurements (i.e., ions) and its high complexity, making data analysis and interpretation laborious. Overcoming these issues and exploiting the potential of metabolomic information promises to unlock valuable insights and optimize monoclonal antibody production.

The state-of-the-art in the Literature relies on the representation of this type of system (i.e., cell cultures) using kinetic models (Kyriakopoulos et al., 2018). These are first-principle nonlinear models which provide a comprehensive understanding of the dynamic behavior of cellular metabolic processes. In these models, complexity is associated to a large number of parameters

that define the cell lines behavior and must be carefully estimated from process data. These parameters are associated with specific model equations to regulate the evolution of a particular component; for this reason, model parameters are typically associated either with a chemical, a physical or a biological phenomenon and embed a strong physical meaning. These model parameters are usually estimated by integrating process data in a fitting procedure that attempts to minimize the error between model predictions and the expected trend. Despite the high complexity, kinetic modeling is the most up-to-date and appropriate approach to characterize the dynamic behavior of mammalian cell cultures in terms of cell growth and metabolism (Young, 2013).

Nevertheless, this procedure has a strong limitation since it does not explicitly consider the available set of metabolomics. In the Literature several models try to integrate -omics data in a classical way by refining the kinetic model, where metabolites act as intermediates in equilibrium and can be used to fit new parameters in the model 'equations (Ahn & Antoniewicz, 2012; Ghorbaniaghdam et al., 2013). However, these works use metabolomics data to directly regress the process variable of the cell culture by integration in a first principle models equation or by application data-driven approaches.

Within this framework, the aim of this work is to present a novel approach for integrating metabolomics data in CHO cell lines first principle models to gain valuable process insights and to explore the intricate relationships between biological system phenomena and cell metabolism. Accordingly, metabolites are linked to first-principles model parameters. This approach facilitates the bridging of metabolites with biological phenomena by leveraging all available data and harnessing the power of data-driven methods to establish parameter-metabolite relationships. At the same time, first-principle modeling is used to elucidate the importance of these parameters in understanding the underlying biological phenomena.

3.2 Thesis workflow

The workflow of this Thesis is organized following six steps as shown in the schematic of Figure 3.1.

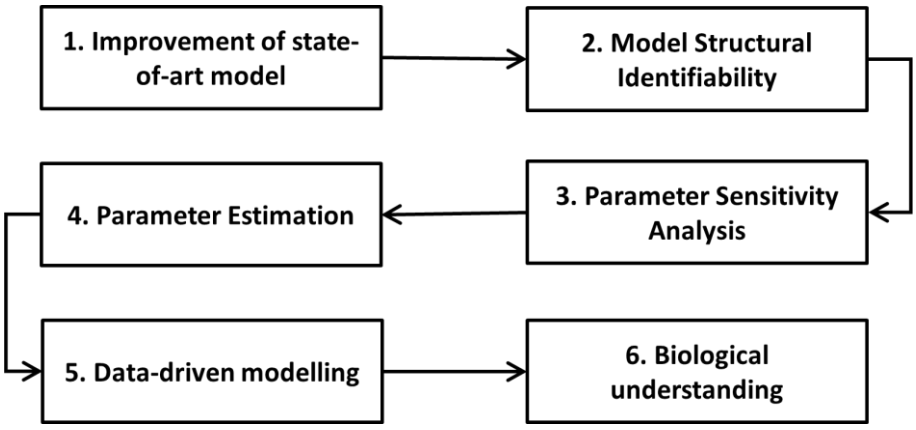


Figure 3.1. *Workflow of the Thesis work.*

1. improvement of the state-of-the-art model. Literature state-of-the-art models for CHO cell lines(Kontoravdi et al., 2010b) are able to discriminate several phenomena (i.e. cell growth, nutrient consumption, etc.). However, the available models are not sufficiently descriptive for the specific case study, so they are improved by adding novel parameters that represent new contributions;
2. model structural identifiability. The structure of the model is mathematically analyzed to determine if all model parameters can be estimated from process data;
3. parameter sensitivity analysis. Due to limited number of experimental points (7) in the course of the cell culture, only 7 parameters can be estimated for each cell line. Sensitivity analysis is performed to rank and identify the 7 parameters that are the most characterizing and the most important to describe the identity of each cell line;
4. parameter estimation. For each cell line, the 7 most characterizing parameters are estimated for each cell line from the process data. The remaining parameters are fixed to a predetermined value, which is not the same for all cell lines, but is specific of the class of cell line defined by the production performance. At this point, the integration of metabolomics data provides an opportunity to explore the biological relationships within the system;
5. data-driven modelling of the relation between the call metabolism and the chemical-physical and biological phenomena occurring in the cell culture. A PLS model is used to relate the cell metabolism in terms of metabolomic dynamics to the most important chemical-physical and biological phenomena occurring in the production of the mAbs, namely the 7 most characterizing parameters of the first-principle model;
6. biological understanding. The results of the data-driven model are explored. The relationships between metabolomics data and model parameters are investigated to determine which metabolites are more influential in determining a specific parameter and thus in controlling a particular biological phenomenon.

3.3 State-of-the-art CHO cell culture model

This Section introduces the state-of-the-art CHO culture model (Kontoravdi et al., 2010b), schematically shown in Figure 3.2.

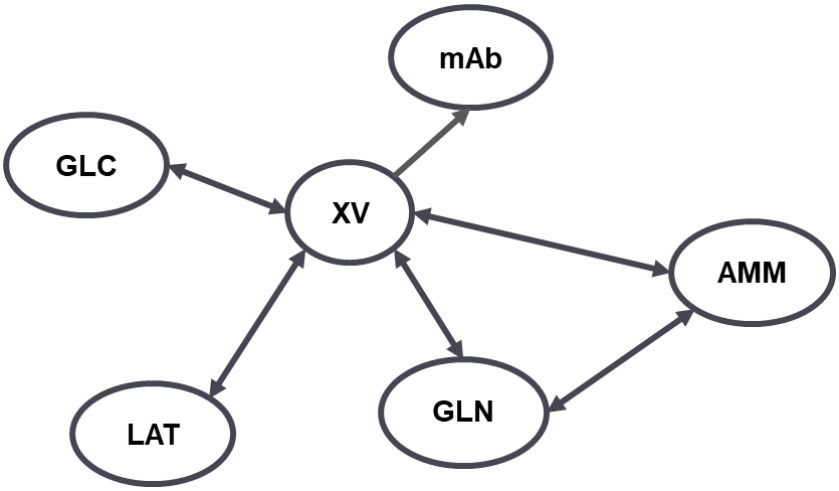


Figure 3.2. *Structure of the original model for the description of CHO cell structure.*

The state-of-the-art model represents the behavior of CHO cell culture by manipulating 8 process variables through a set of differential equations.

- Volume (V):

$$\frac{dV}{dt} = F_{IN} - F_{OUT} \quad (3.1)$$

where F_{IN} and F_{OUT} are respectively the inlet and outlet flowrate;

- Viable Cell Concentration (VCC or X_V)

$$\frac{dX_V}{dt} = (\mu - \mu_D)X_V - \frac{F_{IN}}{V}X_V \quad (3.2)$$

$$\mu = \mu_{max}\left(\frac{C_{GLC}}{K_{GLC}+C_{GLC}}\right)\left(\frac{KI_{amm}}{KI_{amm}+C_{amm}}\right) \quad (3.3)$$

$$\mu_D = \mu_{D,max}\left(\frac{C_{amm}^2}{C_{amm}^2+K_{D,amm}^2}\right) \quad (3.4)$$

where μ_{max} and $\mu_{D,max}$ are respectively the maximum growth and death velocity for the cell culture. K_{GLC} is the parameter that control rate of cell growth due to Glucose consumption and KI_{amm} is the parameter that control Ammonia inhibition in the cell culture. Finally $K_{d,amm}$ is the parameter that control rate of cell death due to Ammonia level;

- Glucose (GLU):

$$\frac{dC_{GLC}}{dt} = \frac{F_{IN}}{V}\left(C_{GLC,IN} - C_{GLC}\right) - \left(\frac{\mu}{Y_{x,GLC}} + m_{GLC}\right)X_V \quad (3.5)$$

where $C_{GLC,IN}$ is the Glucose concentration in the feed and C_{GLC} is the Glucose concentration in the fed batch system. $Y_{x,GLC}$ is the yield parameter that control biomass growth due to Glucose consumption and m_{GLC} is the Glucose maintenance parameter;

- Product Titer (mAb):

$$\frac{dC_{mAb}}{dt} = -\frac{F_{OUT}}{V} C_{mAb} + \left[Y_{mAbglc} \left(\frac{\mu}{Y_{x,GLC}} + m_{GLC} \right) \right] X_V \quad (3.6)$$

where C_{mAb} is the Product Titer and Y_{mAbglc} is the yield parameter that control antibody production due to Glucose consumption;

- Lactate (LAT):

$$\frac{dC_{LAT}}{dt} = -\frac{F_{IN}}{V} C_{LAT} + \left[Y_{latglc} \left(\frac{\mu}{Y_{x,GLC}} + m_{GLC} \right) \right] X_V \quad (3.7)$$

where C_{LAT} is the Lactate concentration and Y_{latglc} is the yield parameter that control Lactate production due to Glucose consumption;

- Glutamine (GLN):

$$\frac{dC_{GLN}}{dt} = -\frac{F_{IN}}{V} C_{GLN} - \left(\frac{\mu}{Y_{x,gln}} + m_{gln} \right) X_V \quad (3.8)$$

where C_{GLN} is the Glutamine concentration in the fed batch system. $Y_{x\text{gln}}$ is the yield parameter that control biomass growth due to Glutamine consumption and m_{GLN} is the Glucose maintenance parameter;

- Ammonia (AMM):

$$\frac{dC_{AMM}}{dt} = -\frac{F_{IN}}{V} C_{AMM} + \left[Y_{ammgln} \left(\frac{\mu}{Y_{x,gln}} + m_{gln} \right) \right] X_V \quad (3.9)$$

Where C_{AMM} is the Ammonia concentration and Y_{ammgln} is the yield parameter that control Ammonia production due to Glutamine consumption;

Unfortunately, the state-of-art kinetic model does not fully capture the complexity of the system under study. In particular, it does not include the role of Glutamate (GLU) as a key nutrient and essential molecule for monoclonal antibody synthesis. To address these limitations, it is essential to update the model to include the contribution of Glutamate and to refine the contribution of other process variables.

3.4 Proposed CHO cell culture model

This Section introduces the proposed CHO culture model, schematically shown in Figure 3.3. As previously explained, the state-of-art model is too simplified to correctly describe the complexity of the biological system under study. As a result, the set of equations governing the model is updated, with particular emphasis on incorporating the new role of Glutamate. This

includes consideration of its function as a nutrient and its ability to regulate Ammonia levels in cell culture.

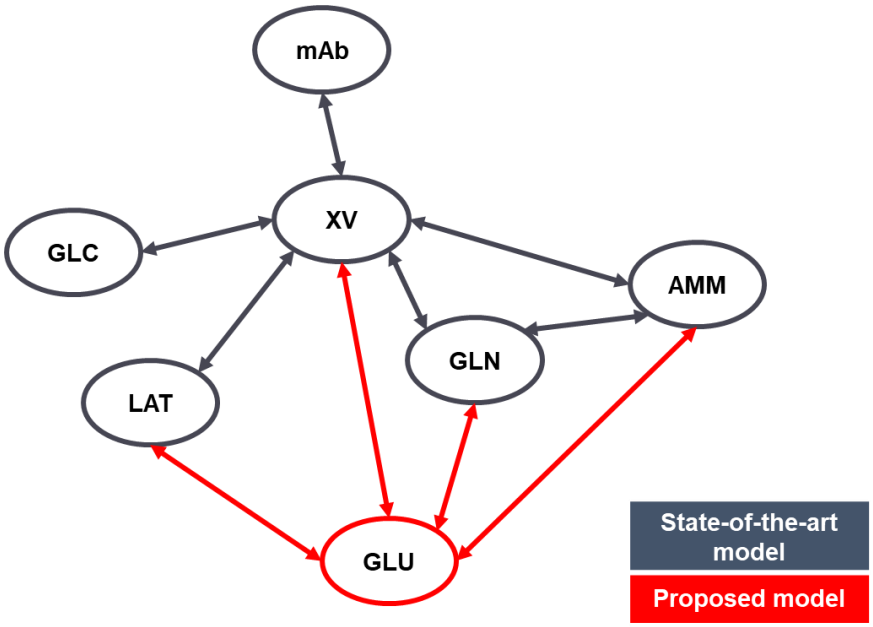


Figure 3.3. *Structure of the proposed model for the description of CHO cell cultures. The red parts are the contributions added to the model.*

In order to improve the modeling of the system under consideration, an investigation of the relationships between various process variables is conducted. In this study, the role of these variables in CHO cell lines is investigated through a comprehensive review of the existing Literature. Furthermore, in the context of the same case study, a previous PCA analysis (Barberi et al., 2022b) revealed interesting relationships between different process variables. These findings provide valuable insights into the interrelationships and dependencies among the variables, contributing to a deeper understanding of the system under study.

After examining the various process variables contributes, the following relationships have been observed in previous Literature and analysis:

- Glutamate is one of the main energy sources in cell cultures (Coulet et al., 2022);
- Glutamine and Glutamate are interchangeable, and their relative amount may be associated to cell needs, such as high energy requirement or molecule synthesis. If no Glutamine is present in the system, it is typically synthetized from Glutamate (Genzel et al., 2005);
- conversion between Glutamate and Glutamine regulates Ammonia level in the cell culture (Schneider et al., 1996);
- cell death rate depends both on both Ammonia and Lactate levels (Krampe & Al-Rubeai, 2010);
- high Lactate concentration along the batch is associated with poor performance (Barberi et al., 2022b);
- high Glutamate concentration tends to increase Ammonia in the system, consequently the cell cultures exhibit worse performance (Barberi et al., 2022b);

- Lactate is generally produced by consumption of Glucose, but it can be also produced from Glutamate during the cell growth phase (Dean & Reddy, 2013);
- Lactate can be used as energy source during the cell growth when Glucose and Glutamate are lacking (Tsao et al., 2005);
- Product Titer and VCC appear to be anticorrelated with Glutamate and Lactate, namely the low is the Glutamate consumption the low will be the final product concentration (Barberi et al., 2022b);

Based on these observations, a novel model structure is proposed. The main changes relate to the role of Glutamate, which is assumed to be a key nutrient component governing the system behavior. The contribution of Lactate is also refined. A new Lactate consumption term is added to account for its conversion back to Glucose, and a new Lactate production term is introduced to account for the conversion of Glutamate to Lactate.

As for the Glutamate, the new proposed model is constructed by introducing two major improvements:

- the relationship between Glutamate and Glutamine is modeled as a reversible first-order kinetics;
- a novel parameter $Y_{glu,X}$ is introduced to describe the behavior of the Glutamate in the first part of the cell culture.

The resulting contribution of the Glutamate, modeled by equations derived from well-established kinetic structures (Botton et al., 2022; Kyriakopoulos et al., 2018), is:

$$\frac{dC_{GLU}}{dt} = \frac{F_{IN}}{V}\left(C_{in,glu} - C_{GLU}\right) - Q_{GLU}X_V + Q_{GLU} + k_1 C_{GLN} - k_2 C_{GLU} C_{AMM} \quad (3.10)$$

$$Q_{GLU} = \frac{\mu}{Y_{x,glu}} + m_{GLU} \quad (3.11)$$

$$Q_{glu,x} = \frac{\mu}{Y_{glu,X}} \quad (3.12)$$

where $C_{in,glu}$ is the concetration of Glutamate in the bolous, k_1 is the kinetic constant regulating conversion of Glutamate to Glutamine, k_2 is the kinetic constant regulating the conversion of Glutamine to Glutamate, and Y_{gluX} is the Glutamate production constant.

Furthermore, the description of Lactate behavior is improved by adding:

- Lactate production due to Glutamate consumption;
- Lactate consumption at low Glucose and high Lactate concentrations
- regulation of cell death ratio by both Ammonia and Lactate concentrations.

Accordingly, the equation that regulates Lactate behavior is refined (Jimenez del Val et al., 2016) and is structured as follows:

$$\frac{dC_{LAT}}{dt} = -\frac{F_{IN}}{V}C_{LAT} + Q_{lat,glc}X_V + Q_{lat,glu}X_V - Q_{lat,cons}X_V \quad (3.13)$$

$$Q_{lat,glu} = Q_{GLU}Y_{lat,glu} \quad (3.14)$$

$$Q_{lat,cons} = \frac{1}{Y_{x,lat}}\left(\frac{C_{lat}}{K_{c,lat}+C_{lat}}\right)\left(\frac{K_{c,glc}}{K_{c,glc}+C_{GLC}}\right) \quad (3.15)$$

$$\mu_D = \mu_{D,max}\left(\frac{C_{amm}^{\alpha_n}}{C_{amm}^{\alpha_n}+K_{D,amm}^{\alpha_n}}\right)\left(\frac{C_{lat}}{C_{lat}+K_{D,lat}}\right) \quad (3.16)$$

where $Y_{lat,glu}$ is the yield of Lactate with respect to Glutamate, Y_{xlat} is the yield of Lactate consumption, $K_{C,lat}$ is the constant regulating the conversion Lactate consumption at low Glucose concentrations and $K_{D,lat}$ is the constant regulating the cell death associated to Lactate. The complete set of equation for the proposed model is reported in Appendix A.

The model proposed and used in this work is finally composed of:

- eight state variables (Table 3.1): $V, C_{glc}, C_{glu}, C_{lat}, C_{amm}, X_V, C_{mAb}, C_{gln}$
- four input conditions (Table 3.1): $F_{IN}, F_{OUT}, C_{glc,in}, C_{glu,in}$
- twenty-five parameters (Table 3.2).

Table 3.1. *Proposed CHO cell model. List of variables used by the model.*

	Variable
V	Fedbatch volume
C_{glc}	Glucose concentration
C_{glu}	Glutamate concentration
C_{lat}	Lactate concentration
C_{amm}	Ammonia concentration
X_V	Viable cell concentration
C_{mAb}	Product titer
C_{gln}	Glutamine concentration
F_{IN}	Inlet flowrate
F_{OUT}	Outlet flowrate
$C_{glc,in}$	Feed Glucose concentration
$C_{glu,in}$	Feed Glutamate concentration

Table 3.2. *Proposed CHO cell model. List of parameters used by the model.*

	Parameter
$K_{c,lat}$	Control factor to Lactate consumption (high Lactate concentration)
$K_{c,glc}$	Control factor to Lactate consumption (low Glucose concentration)
Y_{ammgln}	Yield of Ammonia production due to Glutamine consumption
Y_{ammglu}	Yield of Ammonia production due to Glutamate consumption
Y_{latglc}	Yield of Lactate production due to Glucose consumption
Y_{latglu}	Yield of Lactate production due to Glutamate consumption
Y_{mAbglc}	Yield of Product formation due to Glucose consumption
Y_{xlat}	Yield of biomass growth due to Lactate consumption
k_1	Glutamate to Glutamine constant
k_2	Glutamine to Glutamate constant
K_{glc}	Glucose contribution to cell growth
K_{glu}	Glutamate contribution to cell growth
KI_{lat}	Lactate contribution to cell inhibition
KI_{amm}	Ammonia contribution to cell inhibition
$K_{d,lat}$	Lactate contribution to cell death
$K_{d,amm}$	Ammonia contribution to cell death
μ_{max}	Maximum cell growth rate
$\mu_{d,max}$	Maximum cell death rate
Y_{xglc}	Yield of biomass growth due to Glucose consumption
m_{glc}	Glucose maintenance factor
Y_{xglu}	Yield of biomass growth due to Glutamate consumption
m_{glu}	Glutamate maintenance factor
Y_{xgln}	Yield of biomass growth due to Glutamine consumption
m_{gln}	Glutamine maintenance factor
Y_{glux}	Yield of Glutamate production due to cell activity

3.5 Proposed model structural identifiability

In this Section, the parameters structural identifiability is evaluated for the proposed model. This is a pre-requisite of system identification and parameter estimation since it refers to the ability to determine the values of the unknown parameters from the available input-output data. Structural identifiability ensures that the estimated parameter values are unique and meaningful, allowing for reliable model analysis, validation, and prediction. To perform the latter analysis, the Structural Identifiability procedures (Villaverde et al., 2016) described in Chapter 1 has been applied.

The Observability-Identifiability matrix (OI) is built and due to the model's complexity, the Lie's derivatives are calculated up to the third degree, because higher order calculation results in an excessive computational cost. The process of building the OI by the calculation of Lie's derivatives and the consequently rank calculation is performed with the STRIKE-GOLDD toolbox (Villaverde et al., 2016). The results of the rank calculation of the OI indicates that only ten parameters out of 25 are identifiable at first glance (Table 3.3Table 3.3.Results ***of the structural identifiability procedure. The table reports the parameters identified before (first analysis) and after (second analysis) the model decomposition.***), while no information on the structural identifiability of the remaining parameters is obtained. For this reason, due to the complexity of the system the model is decomposed to reduce the computational time of the

analysis. The decomposition is performed using the MEIGO toolbox (Egea et al., 2014). Additionally, parameters already identified in the first analysis are not considered for further investigation, resulting in fifteen parameters to be studied in the second structural identifiability run. Results of this new structural identifiability run show all parameters as structurally identifiable (Table 3.3).This analysis shows that all the parameters can be potentially estimated from data. However, due to the reduced number of available experimental points (7) for a single cell line, the values of all model parameters cannot be estimated for each cell line. In fact, experimental data for all process variables ($\underline{\mathbf{X}}_P$) are measured in 7 time points. Furthermore, due to model's complexity, the parameter fitting procedure must be carried out considering one variable at a time (as detailed in Section 3.8). Thus, only $(7-1)$ degrees of freedom are available for estimating 6 parameters. In addition to that, an extra parameter (Y_{mAbglc}) can be estimated directly from product titer profile, because it is the only parameter controlling the product titer mass balance.

In this Thesis, all process variables (except the product titer) are used to estimate the value of 6 parameters. After that, product titer profile is used to fit the value of Y_{mAbglc}.

Consequently, the 25 parameters are divided in two groups:

- subset of seven characterizing parameters ($6 + Y_{mAbglc}$) that is selected to guide the behavior of the system and undergo accurate estimation;
- subset of less important parameters whose value is not estimated for each cell line, and they are kept fixed at an average value (between cell lines).

The subset of the seven characterizing parameters is found by investigating which parameters are the most influent in changing the model's response (Sensitivity analysis; Section 3.6).

Table 3.3.*Results of the structural identifiability procedure. The table reports the parameters identified before (first analysis) and after (second analysis) the model decomposition.*

Parameter	Identifiable at first analysis	Identifiable at second analysis
$K_{c,lat}$	•	
Y_{ammgln}	•	
Y_{ammglu}	•	
Y_{latglc}	•	
Y_{latglu}	•	
Y_{mAbglc}	•	
Y_{xlat}	•	
k_1	•	
k_2	•	
K_{glc}		•
K_{glu}		•
KI_{lat}		•
KI_{amm}		•
$K_{d,lat}$		•
$K_{d,amm}$		•
α		•
μ_{max}		•
$\mu_{d,max}$		•
Y_{xglc}		•
m_{glc}		•
Y_{xglu}		•
m_{glu}		•
Y_{xgln}		•
m_{gln}		•

3.6 Parameters sensitivity analysis to identify the most characterizing parameters

In this Section, the parameters sensitivity analysis is performed, to understand the sensitivity of the model's responses to changes in the parameters. This serves to prioritize the parameters that have the larger impact on the model's responses.

The sensitivity analysis is performed using two different techniques: *i*) EET analysis (Saltelli et al., 2008a) and *ii*) VBSA (Saltelli et al., 2010). These two types of analysis are characterized by different advantages and disadvantages:

- VBSA is able to analyze the whole parameter space. However, in the proposed biological model the complex feeding strategy and the large number of parameters may cause large numerical instability;
- EET is simpler and faster but does not ensure to investigate the whole parameter space.

In summary, VBSA is a more comprehensive approach with respect to EET, but it lacks in robustness when applied to complex systems. Therefore, to enhance VBSA applicability, it is integrated with the EET.

3.6.1 Simplified sensitivity analysis

In this Section, two methodologies for sensitivity analysis, EET sensitivity indices and the VBSA resulting Sobol's indexes, are compared to clarify which one is the best for the system under study and to understand if EET analysis can be used as a substitute for VBSA.

To compare the two methodologies, the two analyses are executed considering a simplified feeding strategy, which facilitates the application of VBSA.

The Elementary effect analysis is applied to the model following the procedure described in Chapter 1. The main considerations for the analysis are:

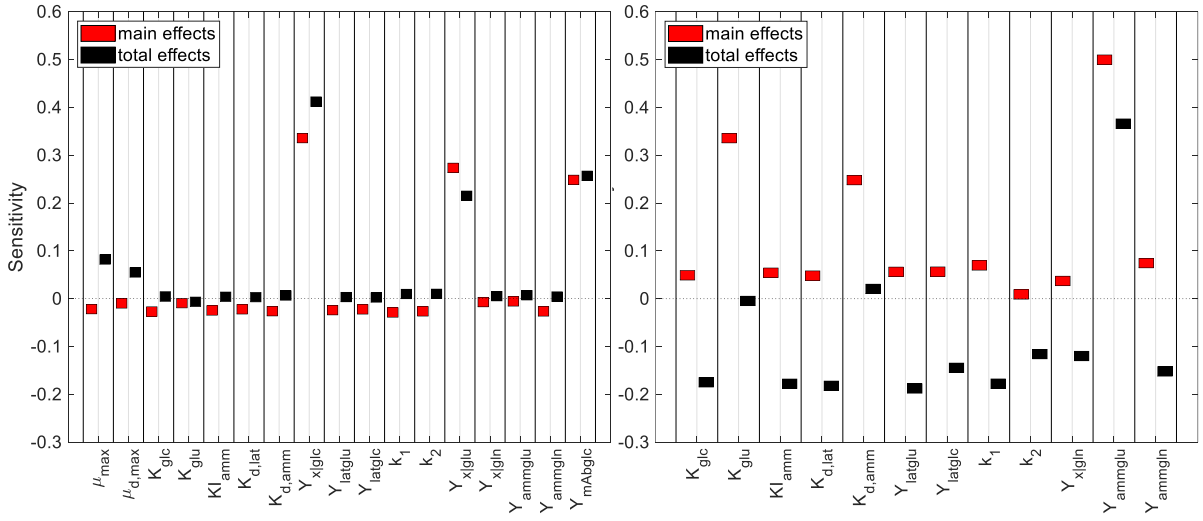
- parameters are free to vary up to 20% around their initial values;
- initial parameter values are obtained by a preliminary fitting;
- parameters are sampled from a uniform distribution;
- the selected model's response is the viable cell concentration;
- $1 \cdot 10^5$ runs are selected in the EET Montecarlo analysis to assess the robustness of the results.

Results of the EET analysis suggests that 5 parameters induce important changes in the model responses and are identified as important parameters. These parameters are: $\mu_{max}, \mu_{dmax}, Y_{Xglc}, Y_{Xglu}, Y_{mAbglc}$.

Once the important parameters have been identified by the EET the the sensitivity index using a Variance-based approach can be calculated. VBSA is used to calculate Sobol's Index (Saltelli et al., 2008b) following the procedure described in Chapter 1. The assumptions considered in the application of VBSA are the same as those of EET to ensure a fair comparison between the two methods.

Results of the Variance Based analysis are reported in Figure 3.4a. Three parameters are identified as important $(Y_{xglc}, Y_{xglu}, Y_{mAbglc})$, having both the main and total effect siginicantly different from zero. Two additional parameters $(\mu_{max}, \mu_{d,max})$ exhibit Sobol's indices (total effect) that are significantly different from zero and negligible main effects. This means that these parameters may have a negligible impact when considered individually (non-significant first order interaction). However, when they interact with other parameters, their overall importance becomes visible as they significantly influence the model's response.

To gain a better understanding of the contributions of parameters with Sobol's index close to zero, another round of VBSA is performed, by fixing the value of the most important parameter identified in the initial analysis. In this way, the variance of the response (i.e. $V(Y)$) becomes solely associated with the few important parameters, allowing for a more in-depth investigation of their behaviors. Following this rationale, the Sobol's indexes resulting from this second variance-based run are presented in Figure 3.4b**.** In this Figure, a new parameter (Y_{ammglu}) demonstrates relative importance in determining the behavior of the cell culture, both its main and total effects are significantly different from zero.



(a) (b)

Figure 3.4. *Results of the VB sensitivity analysis with the simplified feeding strategy: (a) procedure applied to 17 parameters, and (b) procedure applied to 12 parameters. The square represents the mean value of the sensitivity index.*

The overall results of the VBSA are summarized in Table 3.4, which presents the outcomes of the first analysis, and identifies the parameter that captures the majority of the response variability, as well as the results of the second analysis, identifying parameters that contribute to a lesser extent.

Comparing the results obtained from the two analyses (Figure 3.5), both methods successfully identify the subset of the most important parameters. Even if some discrepancies arise in identifying the least important parameters, these do not pose a significant issue in this specific case, because only the most relevant factors will be considered in the estimation phase.

Table 3.4. *Results of the VBSA sensitivity analysis applied to the simplified feed.*

Important Parameters	Probably important Parameters
$\mu_{max}, \mu_{dmax}, Y_{Xglc}, Y_{Xglu}, Y_{mAbglc}$	Y_{ammglu}

In conclusion, this Section demonstrates that in the studied biological system, the EET produces identical results to the VBSA in identifying the most important parameters. Although this demonstration is conducted on a simplified feeding schedule, we can reasonably assume its general validity, even in more complex scenarios. For this reason, EET will be used in this Thesis as sensitivity analysis methods, as it is simper, stabler, and faster.

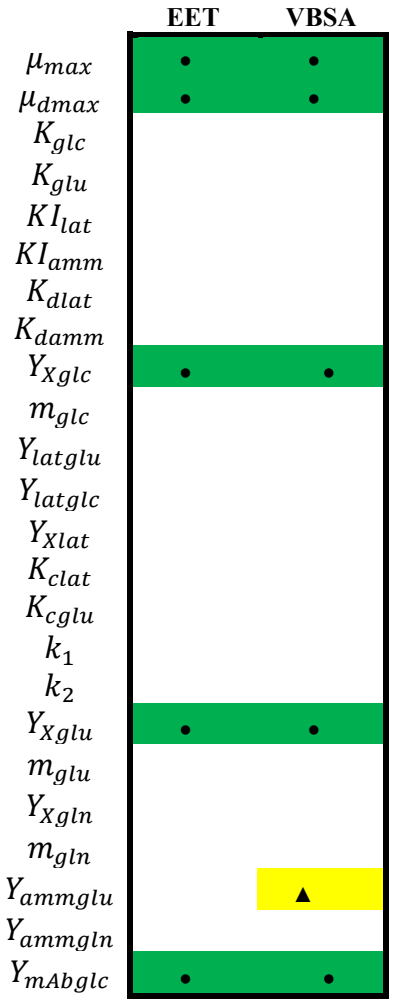


Figure 3.5. *Results of different sensitivity analysis methodologies. Green (●) - most important parameters; yellow (▲) - parameters supposed to be important.*

3.6.2 Identification of most characterizing parameters by EET analysis

In this Section, a complete sensitivity analysis is conducted by considering:

- experimental feeding strategy: the simplified feeding strategy is replaced with the one used in the experiments. As a result, the mean values of model parameters need to be updated;
- multiple responses: the sensitivity analysis is calculated for multiple responses to gain a comprehensive understanding of the influence of parameters on all the process variables;
- dynamic behavior: the sensitivity index of each parameter is calculated in multiple time points because the importance of parameters is expected to vary along the entire life of the cell culture (Kontoravdi, 2006).

These additional complexities pose significant challenges in obtaining accurate results from the sensitivity analysis. Accordingly, to enhance the results, the following assumptions are considered:

- the simulated response is constrained to be non-negative because all state variables represent physical entities that cannot have negative values. By enforcing this mathematical

constraint, all the responses are guaranteed to remain greater than zero throughout the integration, ensuring physically meaningful outputs;

- numerical errors associated with stiff equations are mitigated by modifying the integration procedure, to expand the parameter space that can be explored. The integration is performed in a piecewise manner between feeding actions, while the effect of daily feed boluses is determined through separated by mass balances;
- parameters can vary by 20% around their nominal values. Further variations beyond this threshold may introduce numerical errors due to inherent mathematical issues associated with the model itself.

Considering these assumptions, the sensitivity analysis is performed considering the same setting as in Section 3.6.1.

Before conducting the EET analysis, the effect of the analysis settings on the sensitivity results is studied, and shows that:

- analysis is robust and stable: a larger number of runs in the Monte Carlo simulations do not result in any significant change in the obtained results;
- type of parameter prior distribution has negligible impact on the outcomes: changing the parameter distribution from uniform to normal does not alter the final results of the analysis;
- parameter range has a negligible effect on results: a smaller range of parameters has no effect on the results of the sensitivity analysis leading to identification of the same important parameters. Note that it is crucial to select an appropriate dimension range to ensure a good exploration of the parameter space.

The results of the sensitivity analysis are presented in Figure 3.6 (product titer) and Figure 3.7 (viable cell concentration), while the outcomes for all other variables are detailed in Appendix B. The analysis reveals that the importance of different parameters undergoes significant changes over time. Some parameters maintain their importance throughout the entire duration of the experiment, such as μ_{max}, which consistently plays an essential role in determining the response of the VCC.

On the other hand, other parameters, such as Y_{xglu}, become influential in determining the response of VCC only during later stages of the cell culture's lifespan. This confirms that the importance of specific parameters may vary depending on the specific time instant considered for the analysis.

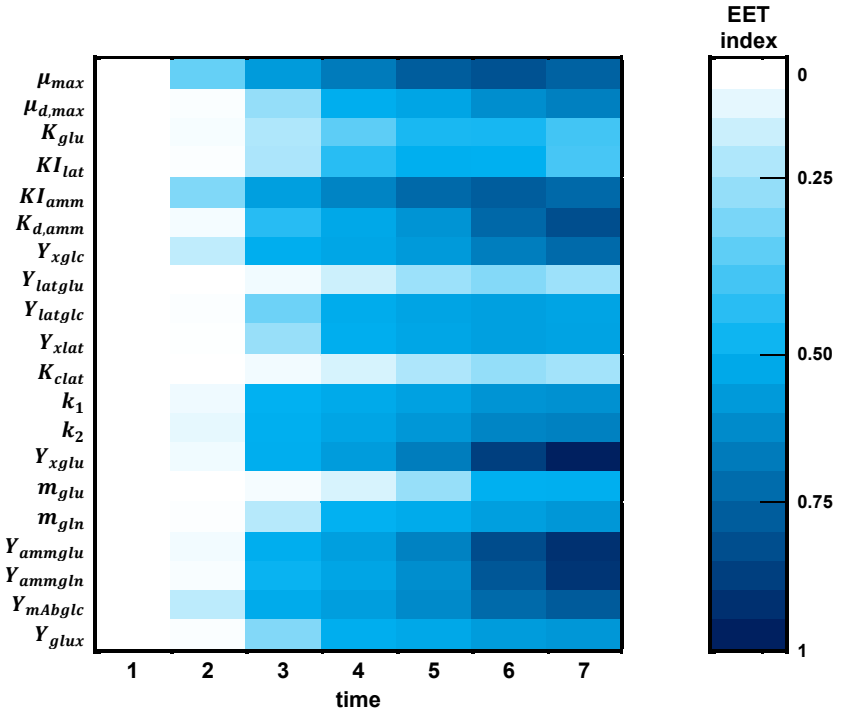


Figure 3.6. *Results of the EET sensitivity analysis for the Product Titer.*

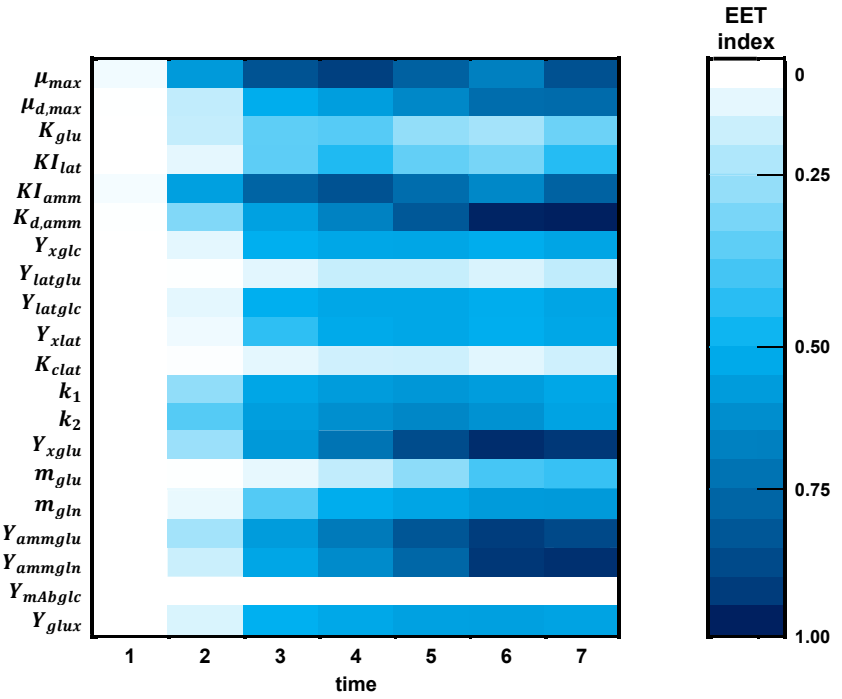


Figure 3.7. *Results of the EET sensitivity analysis for the viable cell concentration.*

The obtained sensitivity EET indices are related to a specific response. For this reason, the important parameters for a specific response along the entire cell culture can be assessed through the cumulative (in time) values of the EET indexes. The selection follows the rationale described below and results are collected in Table 3.5.

- multiple variable control: if a parameter is found to control more than two variables, it is considered as highly important;
- sole variable control: if a parameter has exclusive control over specific variables, it is considered as highly important since it is critical for manipulating those particular aspects of the system;
- subset exclusivity: if a subset of important parameters is chosen to control other variables according to one of the two previous rules, an additional significant parameter has to be identified specifically for that particular variable.

Following latter indications, a subset of seven (6 + Y_{mAbglc}) very important parameters is identified: $\mu_{max}, KI_{amm}, Y_{Xglu}, Y_{latglc}, Y_{gluX}, Y_{Xglc}, Y_{mAbglc}$. These parameters (i.e. characterizing parameters) are the main contributors to the overall variability of the system and control the main biological phenomena. Nutrient consumption is related to Y_{Xglc} and Y_{Xglu}, cell growth to μ_{max}, while Ammonia and Lactate inhibition are related to KI_{amm} and Y_{latglc}. Finally, the formation of monoclonal antibody associated with cell activity is regulated by Y_{mAbglc}.

The remaining parameters are identified as less important for the proposed model as their variation does not induce significant changes in the response variables.

Table 3.5. *Important parameters for all the culture variables: (●) indicates very important parameters in determining the value for the associated variable according to the cumulative EET index.*

	VCC	**Glucose**	**Lactate**	**Glutamate**	**Product Titer**
μ_{max}	●	●		●	●
Y_{Xglu}	●	●			●
K_{damm}	●				
Y_{ammglu}	●				●
KI_{amm}	●	●		●	●
Y_{ammgln}	●				
μ_{dmax}					
k_2					
k_1					
m_{gln}					
Y_{gluX}				●	
Y_{Xglc}		●	●		
Y_{latglc}			●		
Y_{Xlat}			●		
KI_{lat}					
K_{glu}					
m_{glu}					
Y_{latglu}					
K_{clat}					
Y_{mAbglc}					●

3.7 Clustering of cell behavior through MPCA and k-means

In this Section, the objective is to identify groups of the cell lines with different productivity behaviors, which can be described in a similar manner and with the same model parameters. This operation is necessary because, as mentioned in previous Sections, only seven characterizing parameters undergo precise estimation for each cell line, while the remaining parameter cannot be individually fitted and are fixed at a value that represents the general behavior of the biological system. This is done to apply fixed parameters across cell lines with similar behavior, avoiding generalizing the whole system variability in a single behavior, which would impose significant limitations and proves inadequate as it compromises the quality of subsequent estimation steps.

To address this issue, the cell lines are divided into different classes and specific values are assigned to the fixed parameters for each cluster, capturing the unique characteristics of the respective sub-family. The process of clustering is accomplished in a reduced (latent) space obtained through MPCA, which effectively summarizes the entire dynamic behavior of cell cultures. The actual clustering is performed through k-means clustering applied on the MPCA scores. Detailed mathematical methods for MPCA and the k-means clustering are explained in Chapter 1.

3.7.1 Modelling cell dynamic behavior through MPCA

In this Section, the cell dynamic behaviors are modeled in a reduced dimensional space for either a better recognition of similarities among cell cultures or the identification of differences. To achieve this objective, MPCA is applied to the process dataset ($\mathbf{X_P}$) so that the correlation between process variables (e.g. Glucose, VCC, mAb, etc.) can be used to study dynamic cell behavior. As the available experimental measurements are performed at seven different time points, MPCA is used to effectively capture the dynamic behavior of the biological system. The number of latent variables is selected based on the minimization of the RMSECV, which in this case is calculated through a Venetian blind cross-validation procedure. The explained variance and the value of the RMSECV for each principal components are presented in Table 3.6. A total of 9 PCs minimize the RMSECV capturing approximately 90% of the overall variability, which is considered quite satisfactory within the context of a biological system. Finally, residuals are checked to ensure normality and absence of trends.

Table 3.6. *MPCA model on cell-culture data: cumulative and per PC explained variance, and RMSECV. The bold underlined character shows the selected number of PCs based on the minimization of the RMSECV.*

PC	Explained Variance [%]	Cumulative explained variance [%]	RMSECV
1	39.7	39.7	0.844
2	15.4	55.1	0.780
3	9.47	64.5	0.731
4	8.16	72.7	0.683
5	4.82	77.5	0.657
6	4.28	81.8	0.644
7	3.07	84.9	0.655
8	2.83	87.7	0.640
9	**2.52**	**90.2**	**0.630**
10	1.73	91.9	0.641

Identification of main cell behaviors is limited to the first two principal directions because they capture a significant part of the overall data variability (55.1%). Their loadings reveal the most relevant correlations between the process variables within the described biological system. Correlations captured by the first two PCs are summarized in Figure 3.8:

- PC1: Viable Cell Concentration and product titer are positively correlated and anticorrelated with Lactate and Glutamate. Specifically, the cell cultures with negative PC1 scores exhibit lower VCC and final product concentration, lower Glutamate consumption, and higher Lactate production;
- PC2: Glucose and Glutamate are anticorrelated with Lactate. Positive PC2 scores have higher levels of Glucose and Glutamate, with lower levels of Lactate. Along this direction, the values of VCC and Product Titer do not exhibit significant changes.

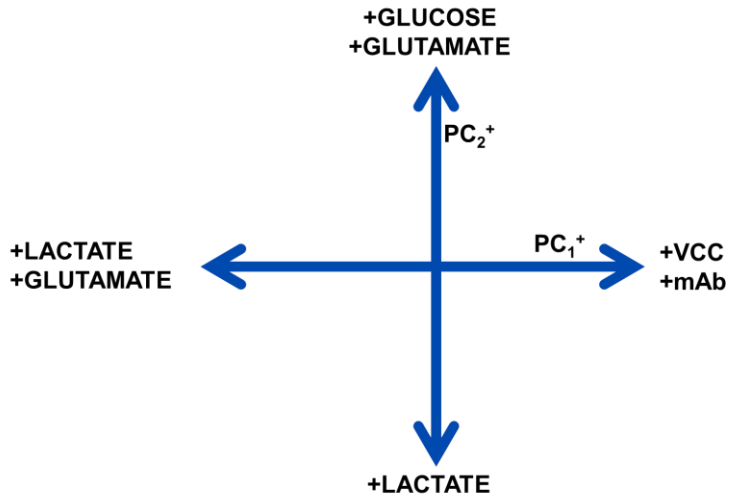


Figure 3.8. *Correlations among main culture variables captured by PC1 and PC2 in the MPCA model on cell culture data.*

3.7.2 Clustering of cell behavior

In this Section, the information obtained by modeling cell-lines behavior according to the experimental data in the reduced latent space is used to cluster the cell-lines different groups. This operation is necessary to distinguish different groups according to productivity behavior that appear in the biological system under study. To carry out the clustering, a k-means algorithm is used. The methodology allows analyzing the score space derived from the previously built MPCA model and dividing the cell lines in 3 groups[1] according to the dynamic behaviors of the variables, which are the most influential on the direction of maximum variability of the data (PC1), which are VCC and product titer, the main indicators of cell line quality and performance. The results of the clustering procedure are presented in Figure 3.9. This means that the three groups distinguish the cell line behavior according to the main dynamics of VCC and product titer. In fact, the central cluster (△) represents batches with a standard productivity behavior of VCC and product titer dynamics. On the other hand, the cluster on the left contains the low-performing batches (○). These batches display high lactate concentration and great inhibition, resulting in lower VCC and product titer values than other cell lines. In the right part of the score space, we find the high-performing batches (×). These batches demonstrate improved consumption of Glutamate and increased resilience to Lactate. Consequently, they yield high values of VCC and product.

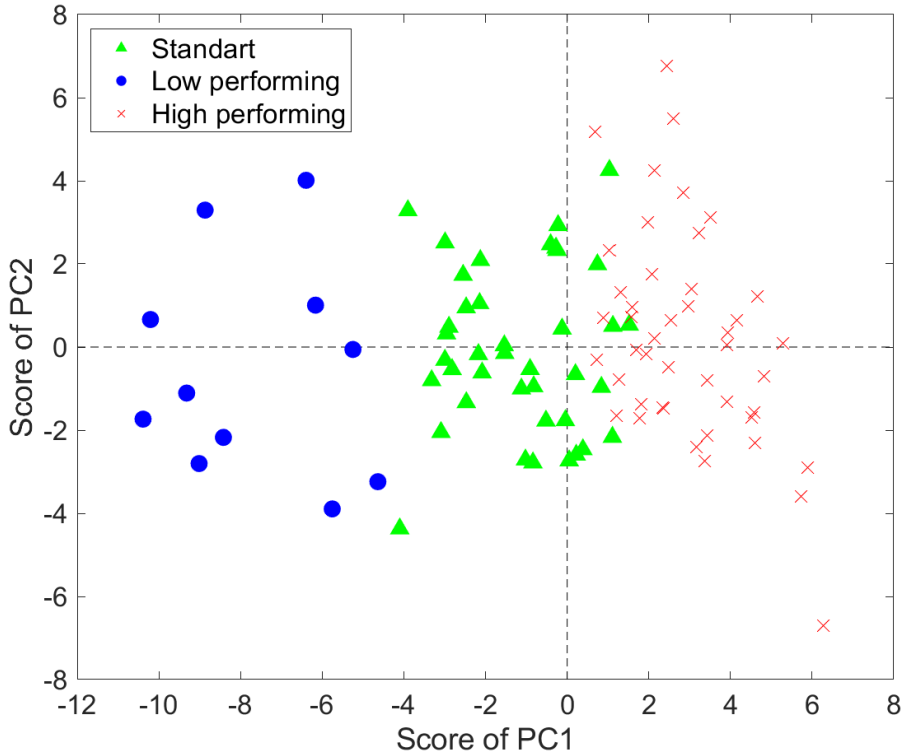


Figure 3.9. *K-means clustering on the score space of the MPCA model on cell culture data.*

[1] Note that a lower number of cluster captures behaviors that are not sufficiently different, while larger number of clusters identify small groups that show only limited differences on the overall behavior.

3.8 Estimation of first-principle model parameters from process data

In this Section, the methodology employed for parameter estimation is explained in detail. This is essential to determine the values of the model parameters for each cell line. As demonstrated in previous sections, the set of parameters has been divided into two classes of importance: seven characterizing parameters, that require to be fitted for each cell line, and a subset of remaining parameters whose values are set as constants.

First of all, to ensure a reliable estimation, it is essential to start from an acceptable initial point, which is identified by retrieving parameter values from the Literature and refining them through a preliminary fitting. These refined values are referred to as adjusted literature-parameter values.

Once the initial set of parameters is determined, the estimation procedure is approached separately for the fixed parameters and the characterizing parameters.

In the determination of the fixed parameters, their values are calculated separately for each of the three clusters identified in the previous section. This approach avoids the estimation average values across all the cell cultures. Instead, it allows determining specific parameters for each group of cells, according to their VCC and titer dynamics, and to characterize their macro-trends within biological system.

For the seven characterizing parameters, their values are individually fitted for each cell line. These parameters are essential for capturing the differences among various lines, while highlighting the specific characteristic behavior of each one. However, the methodology for parameter estimation is inherently complex because the limitations of the biological model, the large number of parameters involved, and the lack of multiple experimental data points make it challenging to obtain accurate estimations. Therefore, we need a systematic approach to perform the estimation in the form of a step-by-step method. The estimation strategies vary depending on the type of parameters being considered (characterizing, fixed or adjusted literature parameters). Two distinct strategies are used in this Thesis:

- fixed parameters*:* for estimating the parameter values: *i)* all cell lines within a single group are included in the parameter estimation process. However, *ii)* due to the limited availability of experimental data points, only seven parameters can be estimated simultaneously. For fixed parameters *iii)* all parameters are set to start by their adjusted-literature value; during the estimation, *iv) v*arious combinations of parameters can be adjusted to refine the estimation gradually, enabling the fitting of all parameters. In the procedure, *v)* all variables are retained to prevent the overfitting of any specific process variable. In addition, at each estimation step *vi)* only few iterations are retained; specifically, the ones that contribute most to the decrease of the residual errors. In fact, the risk is that the optimization solver, in its attempt to minimize the objective function, may find values for the parameters that differ significantly from the literature ones;

- adjusted literature parameters: the fitting procedure of the adjusted literature parameters is analogous to that of fitting the fixed parameters. The only differences are related to the fact that: *i)* the starting point for estimation are literature parameters value and *ii)* a single cell line is used to estimate the adjusted literature parameters. Thanks to point *ii*, a large number of iterations can be carried out in the estimation procedure because of the limited risk of overfitting;
- seven characterizing parameters*:* for estimating the parameter values *i)* each cell line is individually considered, resulting in 96 sets of seven parameters specific for each line (**$\mathbf{X_{EP}}$**). For the characterizing parameters *ii)* only seven parameters need to be fitted, while the remaining parameters are held constant at the fixed values previously determined. The initial value of the seven parameters are set to the respective adjusted literature value. The fitting process *iii)* begins with VCC, which plays a central role in the model (Figure 3.2). Once VCC is fitted, the procedure continues by *iv)* adding Glucose contribution and fitting VCC and Glucose together. After that, *v)* this strategy is repeated for all the process variables, adding them one at a time (in order: Glucose, Lactate, Glutamate, Glutamine and Ammonia). Finally, *vi)* Product Titer is incorporated, and the product yield is estimated.

3.8.1 Identification of literature parameters

In this Section, the values of parameters are retrieved from the relevant Literature review to have a reliable starting point for parameter estimation procedure. The parameter values obtained from literature are represented in Table 3.7. Note that the values of some parameters (i.e., K_{cglc}, k_1, k_2, Y_{Xglu}, Y_{mAbglc}, Y_{gluX}) cannot be found in the Literature. For this reason, their initial guesses are hypothesized based on similar parameters; for example, m_{glu} represents the maintenance factor for Glutamate. Since its value is not available in the literature, we assume, as a reasonable starting point, that its order of magnitude is similar to the one of Glucose maintenance coefficient, m_{glc}.

Table 3.7. *Model parameters from the Literature: value, unit of measure, and reference.*

Parameters	Value	Unit	Ref.
μ_{max}	$2.9 \cdot 10^{-2}$	$[h^{-1}]$	(Xing et al., 2010)
μ_{dmax}	$1.6 \cdot 10^{-2}$	$[h^{-1}]$	(Xing et al., 2010)
K_{glc}	$1.5 \cdot 10^{-2}$	$[g_{glu}/L]$	(Xing et al., 2010)
K_{glu}	$4.7 \cdot 10^{-2}$	$[mmol/L]$	(Xing et al., 2010)
KI_{lat}	$3.9 \cdot 10^{3}$	$[mg_{lat}/L]$	(Xing et al., 2010)
KI_{amm}	6.5	$[mmol/L]$	(Xing et al., 2010)
K_{dlat}	$4.1 \cdot 10^{3}$	$[mg_{lat}/L]$	(Xing et al., 2010)
K_{damm}	6.5	$[mmol/L]$	(Xing et al., 2010)
Y_{Xglc}	1.69	$[10^{11} cells/mol]$	(Xing et al., 2010)
m_{glc}	$1.2 \cdot 10^{-11}$	$[g_{glc}/(cell \cdot h)]$	(Xing et al., 2010)
Y_{latglu}	$1.3 \cdot 10^{2}$	$[mg_{lat}/\ mmol_{glu}]$	derived from Y_{latglc}
Y_{latglc}	$7.3 \cdot 10^{2}$	$[mg_{lat}\ /g_{glc}]$	(Jimenez del Val et al., 2016)
Y_{Xlat}	$0.3 \cdot 10^{9}$	$[cell/mg_{lat}]$	(Jimenez del Val et al., 2016)
K_{clat}	$1.2 \cdot 10^{2}$	$[mg_{lat}/L]$	(Jimenez del Val et al., 2016)
K_{cglc}	-	-	-
k_1	-	-	-
k_2	-	-	-
Y_{Xglu}	-	-	-
m_{glu}	$1 \cdot 10^{-12}$	$[mmol_{glu}/(cell \cdot h)]$	derived from m_{glc}
Y_{Xgln}	9.74	$[10^{11} cells/mol]$	(Xing et al., 2010)
m_{gln}	$1 \cdot 10^{-12}$	$[mmol_{glu}/(cell \cdot h)]$	derived from m_{glc}
Y_{ammglu}	$4.5 \cdot 10^{-1}$	$[mmol_{amm}/mmol_{glu}]$	derived from Y_{ammgln}
Y_{ammgln}	$4.5 \cdot 10^{-1}$	$[mmol_{amm}/mmol_{gln}]$	(Kontoravdi et al., 2010b)
Y_{mAbglc}	-	-	-
Y_{gluX}	-	-	-

3.8.2 Adjusted parameter values and comparison with respect to the literature data

In this Section, the Literature parameter values are adjusted so that the proposed model is able to adapt to the cell cultures under study and to better fit the process data. To achieve this, the behavior of a selected cell culture is carefully fitted using the literature parameters as a starting point. The choice of the refence group according to productivity behavior is based on the results of the k-means analysis. The selected cell line is the one closest to the centroid of the standard cell cluster, which is cell line #8. This approach is adopted because it is supposed that this cell culture represents an average point among all the cell cultures, making the retrieved parameters a suitable initial point for all subsequent estimations on our biological system.

The literature values, the results of the preliminary fitting are reported in Table 3.8. In the Table, the variability of each parameter is derived from Literature:(Jang & Barford, 2000; Jimenez del Val et al., 2016; Kontoravdi et al., 2010b; Pörtner & Schäfer, 1996; Xing et al., 2010)

Table 3.8. *Fitted parameter for cell line #8. The value can be used as initial value in the following fitting estimations.*

Parameter	Initial Value	Fitted Value	Literature variability
μ_{max}	-	$1.1 \cdot 10^{-1}$	0.01-1
μ_{dmax}	-	$5.2 \cdot 10^{-1}$	0.01-1
K_{glc}	$1.5 \cdot 10^{-2}$	0	0.01-1.00
K_{glu}	$4.7 \cdot 10^{-2}$	$2.0 \cdot 10^{-1}$	N/A
KI_{lat}	$3.9 \cdot 10^{3}$	$2.8 \cdot 10^{4}$	$100 - 1 \cdot 10^{5}$
KI_{amm}	6.5	$2.3 \cdot 10^{-1}$	1.0-20.0
K_{dlat}	$4.1 \cdot 10^{3}$	0	N/A
K_{damm}	6.5	23	1.00-20.0
Y_{Xglc}	-	1.2	0.1-100
m_{glc}	$1.2 \cdot 10^{-11}$	0	$0 - 1 \cdot 10^{-3}$
Y_{latglu}	$1.3 \cdot 10^{2}$	$8.9 \cdot 10^{2}$	N/A
Y_{latglc}	$7.3 \cdot 10^{2}$	$1.0 \cdot 10^{3}$	N/A
Y_{Xlat}	$0.3 \cdot 10^{9}$	$7.9 \cdot 10^{-2}$	N/A
K_{clat}	$1.2 \cdot 10^{2}$	$1.7 \cdot 10^{2}$	N/A
K_{cglc}	-	$1.0 \cdot 10^{9}$	N/A
k_1	-	$6.4 \cdot 10^{-2}$	N/A
k_2	-	$6.5 \cdot 10^{-3}$	N/A
Y_{Xglu}	-	10	N/A
m_{glu}	$1 \cdot 10^{-12}$	$4.3 \cdot 10^{-5}$	$0 - 1 \cdot 10^{-3}$
Y_{Xgln}	-	$1.0 \cdot 10^{15}$	N/A
m_{gln}	$1 \cdot 10^{-12}$	$1.4 \cdot 10^{-3}$	$0 - 1 \cdot 10^{-3}$
Y_{ammglu}	-	1.2	N/A
Y_{ammgln}	$4.5 \cdot 10^{-1}$	$9.4 \cdot 10^{-1}$	N/A
Y_{mAbglc}	-	68	N/A
Y_{gluX}	-	1.1	N/A

The comparison between adjusted literature parameters and original literature parameters can provide valuable insights on the behavior of the considered cell culture:

- the fitted value of K_{glc} is equal to zero. The parameter represents the Glucose concentration below which the rate of growth of the cell culture is halved. It is likely that the cell cultures are being supplied with an excess of Glucose, consequently, regardless of the specific amount of nutrient provided, the rate of growth is not directly dependent on Glucose consumption. However, other factors associated with Glucose consumption (e.g., Lactate production) can indirectly limit the growth of the cell culture;
- the parameter K_{dlat} represents the rate of cell death in relation to Lactate concentration. The fitted value of $K_{dlat}$is zero, indicating that rate of cell death is independent on the quantity of Lactate in the cell culture;
- K_{cglc} is the constant that regulates Lactate consumption at lower glucose concentration. The fitted value for this parameter suggests that this contribution can be neglected. This means that, Lactate consumption can be described solely based on its concentration.

3.8.3 Fixed parameter values estimation

In this Section, the results of the estimation for fixed parameter are presented for each group and compared. The estimation procedure for the fixed parameters follows the indication previously described. Fitted parameters are reported in Table 3.9**.** No significant changes with respect to the adjusted literature values are obtained for cluster of standard productivity cell line.

Table 3.9. *Value of fixed parameters for the three groups of cell cultures. Parameters underline exhibit greater changes respect to adjusted literature parameters.*

Parameter	Adjusted literature value	Standard productivity cell lines	High performing cell lines	Low performing cell lines
μ_{dmax}	$5.167 \cdot 10^{-1}$	$5.012 \cdot 10^{-1}$	$2.443 \cdot 10^{-1}$	$2.774 \cdot 10^{-1}$
K_{glc}	0	0	0	0
K_{glu}	$1.983 \cdot 10^{-1}$	$7.721 \cdot 10^{-1}$	$\mathbf{4.680 \cdot 10^{1}}$	$1.490 \cdot 10^{1}$
KI_{lat}	$2.819 \cdot 10^{4}$	$4.554 \cdot 10^{4}$	$\mathbf{1.864 \cdot 10^{7}}$	$6.322 \cdot 10^{4}$
K_{dlat}	0	0	0	$\mathbf{9.482 \cdot 10^{1}}$
K_{damm}	$2.255 \cdot 10^{1}$	$2.540 \cdot 10^{1}$	$1.784 \cdot 10^{1}$	$2.456 \cdot 10^{1}$
m_{glc}	0	$\mathbf{1.814 \cdot 10^{-3}}$	$6.700 \cdot 10^{-4}$	$6.105 \cdot 10^{-3}$
Y_{latglu}	$8.906 \cdot 10^{2}$	$4.242 \cdot 10^{3}$	$\mathbf{3.658 \cdot 10^{4}}$	$2.591 \cdot 10^{4}$
Y_{Xlat}	$7.906 \cdot 10^{-2}$	$8.917 \cdot 10^{-2}$	$9.790 \cdot 10^{-3}$	$\mathbf{4.050 \cdot 10^{1}}$
K_{clat}	$1.691 \cdot 10^{2}$	$1.661 \cdot 10^{2}$	$3.341 \cdot 10^{2}$	$1.433 \cdot 10^{4}$
K_{cglc}	$1.000 \cdot 10^{15}$	$1.000 \cdot 10^{15}$	$1.000 \cdot 10^{15}$	$1.000 \cdot 10^{15}$
k_1	$6.388 \cdot 10^{-2}$	$1.041 \cdot 10^{-1}$	$3.228 \cdot 10^{-2}$	$\mathbf{7.283}$
k_2	$6.487 \cdot 10^{-3}$	$1.760 \cdot 10^{-2}$	$5.489 \cdot 10^{-3}$	$1.992 \cdot 10^{-1}$
m_{glu}	$4.320 \cdot 10^{-5}$	$7.428 \cdot 10^{-5}$	$4.558 \cdot 10^{-7}$	$9.090 \cdot 10^{-6}$
Y_{Xgln}	$1.000 \cdot 10^{15}$	$1.000 \cdot 10^{15}$	$1.000 \cdot 10^{15}$	$1.000 \cdot 10^{15}$
m_{gln}	$1.425 \cdot 10^{-3}$	$3.890 \cdot 10^{-3}$	$1.921 \cdot 10^{-3}$	$2.553 \cdot 10^{-3}$
Y_{ammglu}	1.170	1.807	$\mathbf{1.055 \cdot 10^{2}}$	$3.652 \cdot 10^{1}$
Y_{ammgln}	$9.417 \cdot 10^{-1}$	1.030	$5.760 \cdot 10^{-1}$	$5.884 \cdot 10^{-2}$

For the high overperforming cell cultures, the most significant changes are related to how the cell culture responds to Lactate, Ammonia, and Glutamate. These batches are characterized by distinct feeding strategy that is less consistent in terms of Glutamate. As a result, K_{glu} has a greater value compared to standard productivity batches, meaning that Glutamate is consumed more efficiently in high performing cell culture. Furthermore, the improved performance of these batches appears to be associated with a great resilience to Lactate. In this cell lines Lactate concentration is similar to standard productivity cultures, however the high value of parameter KI_{lat} indicates that cells are inhibited only at very high Lactate level. Additionally, the levels of Lactate and Ammonia remain consistent. To counterbalance the low Glutamate consumption, the yield constant associated with Glutamate (Y_{latglu}, Y_{ammglu}) has higher values Y_{latglu}, Y_{ammglu}.

Finally, comparing the fixed parameter values of low productivity cell cultures, the most relevant differences are related to parameter: K_{dlat}. This parameter influences the death of the cell culture in response to Lactate concentration. In standard productivity and high performing

cultures, the fitted value is equal to zero (not relevant), but in low performing batches it becomes significantly different from zero. It seems that for this cell cultures Lactate has not only a contribution to the inhibition of the cell growth, but it contributes also to the death of the cell culture, and this is the main reason for which both VCC and Product Titer are low.

3.8.4 Cell culture most characterizing parameters estimation

In this Section, the estimation of the subset of the seven characterizing parameters is performed for each cell line. In Figure 3.10 is reported an example of fitting proposed for the cell culture #47 (standard productivity cell line). The model effectively captures the decreasing profile of Glucose consumption (Figure 3.10a) and the trend of VCC, both during first half of cell culture and at variable peak (Figure 3.10c). The model successfully predicts the maximum of the Lactate profile; however, it struggles to accurately identify the Lactate trend in the later time points (Figure 3.10b). On the other hand, the model performs very well in capturing the production trend of the monoclonal antibody (Figure 3.10d).

The fitting performance is presented in Table 3.10, which provides the average (between all time points) fitting error in terms of MRE and the MEA scaled to the maximum response for each process variable. Both VCC and Product Titer show a highly accurate fit, with low error. The MRE of lactate is higher, primarily due to the low fitting performance in the final time instants of the cell culture. However, it is important to consider that the MRE value can be influenced by very low values. Therefore, it is crucial to double-check and verify the goodness of fitting using the MEA value, which certify the accuracy of the fit for Lactate.

Table 3.10. *Statistics of the fitting for cell culture n° 47. MRE and MEA/max(y) for each process variable.*

	VCC	**Product Titer**	**Glucose**	**Lactate**
$\boldsymbol{MRE}$	11.5%	16.3%	24.8%	105%
$\boldsymbol{MEA\,/\,max(Y)}$	6.65%	2.75%	8.46%	10.6%

Given that the estimation results are obtained by fitting only seven parameters while keeping the fixed parameters constant, the overall fitting results can be considered satisfactory.

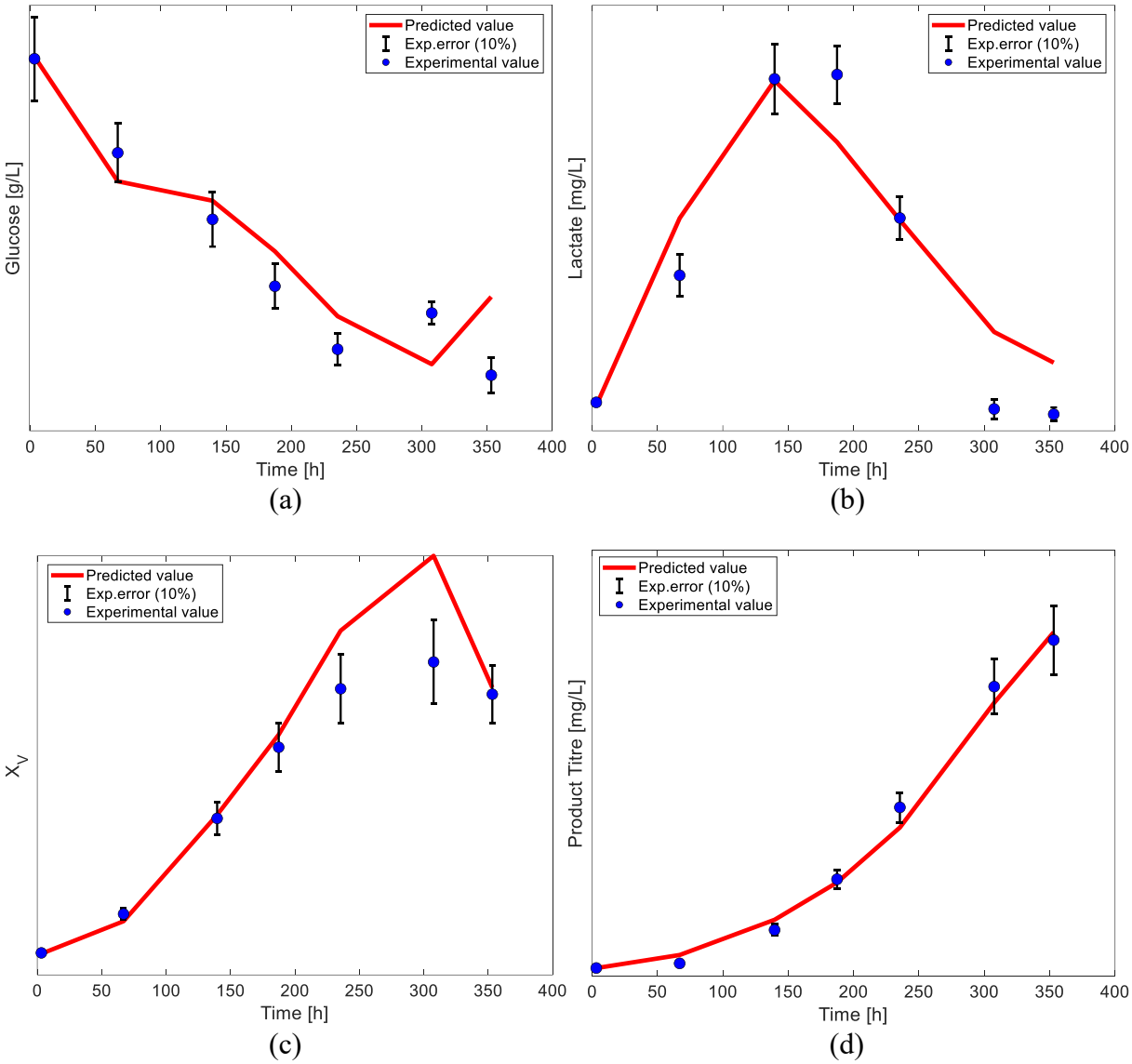


Figure 3.10. *Results of the parameter estimation and process variables fitted procedure for cell line #47. In the figure different behaviors are described: Glucose (a), Lactate (b), VCC (c) and the final product (d).*

Chapter 4

Relating cellular metabolism to chemical, physical and biological phenomena in CHO cultures

In this Chapter, metabolomics data and process information derived from the first-principle model of Chapter 3 are integrated to extract valuable insight on how cell metabolism relate to the chemical, physical and biological phenomena occurring into the cell culture. This is achieved by building multivariate regression models that relate the estimated first-principles model parameters and the metabolomics data.

4.1 Relating cell line metabolism to chemical-physical and biological phenomena using PLS modelling

This Section presents the construction of the PLS model to investigate the relationships between physical, chemical, and biological phenomena of cell culture and cell metabolism. This investigation is performed by a mathematical regression model that relates metabolomics data to first-principles model parameters. In this Thesis, seven PLS models are considered, each one relating one of the 7 *highly characterizing* parameters to the metabolomics data. The models are built on the batchwise unfolded metabolomics dataset ($X = \mathbf{X_{IC}}\ [96 \times 27522]$) and regress the value of first-principle model parameter ($Y = \mathbf{X_{EP}}\ [96 \times 7]$).In each model, $\mathbf{X_{IC}}$ is mean-centered and Pareto-scaled, while $\mathbf{X_{EP}}$ are auto scaled.

Table 4.1. *PLS model calibration. Statistics of all models built for each first-principle parameter.*

	μ_{max}	Y_{xglc}	Y_{mAbglc}	Y_{xglu}	Y_{glux}	KI_{amm}	Y_{latglc}
LV	5	2	7	6	4	7	4
R_y^2	0.983	0.722	0.969	0.99	0.953	0.989	0.972

Calibration results for each of the seven PLS models are presented in Table 4.1, in all cases, the number of latent variables is selected by minimizing the RMSECV through venetian blind cross-validation.

To obtain these results, before constructing the PLS model, $\mathbf{X}_{IC}$ undergoes a data enhancement process to handle missing data, as described in Section 4.1.1. Then, $\mathbf{X}_{IC}$ is used to build a multiway PLS model to regress the value of X_EP.
Additionally, the quality of PLS models is improved through an iterative process that focuses on retaining only the most informative ions to predict the values of the first-principles model parameters (Section 4.1.2). This iterative process ensures a strong relationship and enables reliable results when investigating the link between cell metabolism and cell biological phenomena.

4.1.1 Address missing data in the ion dataset for improved analysis

In this Section, the procedure applied for preprocessing metabolomics dataset is explained. Pre-processing is essential to handle potential missing information and to ensure the reliability and validity of the model. In the original metabolomics dataset a replacement technique for ions with missing intensities is used to ensure a reliable analysis (Barberi et al., 2022a). This method involves imputing the missing values by calculating the weighted average intensity of 15 metabolites that have intensity profiles similar to the metabolite of interest. However, to prevent the introduction of artifacts due to an excessive number of missing values, additional preprocessing is considered. This involves comparing the missing data values in the two replication matrices; the procedure (Figure 4.1) is built on a three-step operation:

1. ions are excluded from the analysis if a number of missing data greater than 20% in both replicates, (Figure 4.1a) is found. This happens when the ion was not identified by the mass spectrometer;
2. if an ion has a missing data count exceeding 20% in only one of the two replicates, the mean of relative difference between the ion intensity of the two replicates is calculated. If the difference is >20%, the ion is excluded from the analysis (Figure 4.1b), because it is considered as poorly repeatable and detectable;
3. ions are removed from the analysis if they have a mean relative difference in the intensity larger than 25% between the two replicates (Figure 4.1c). For those ions, the same m/z does not provide consistent and repeatable information within the two replicates.

Following this rationale, the 4.44% of the original ions are removed.

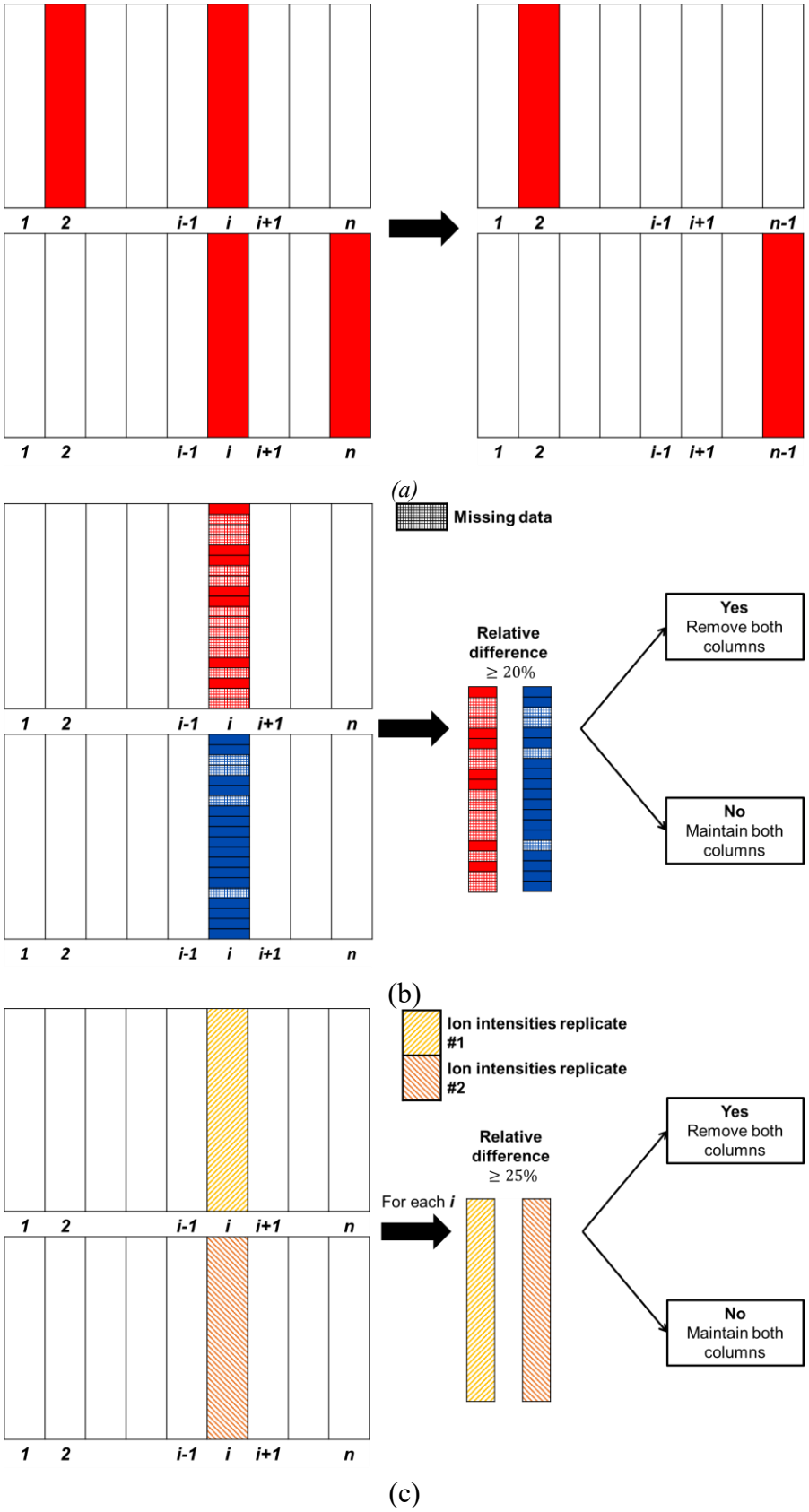


Figure 4.1. *Preprocessing procedure: a) ions with excessive number of missing values in both replicates are removed; b) ions with an excessive difference in missing data between the two replicates are removed; c) ions with large intensity differences between the replicates are removed.*

4.1.2 Improve model performance by selecting the most informative ion lines

This Section shows how the PLS model built on the metabolomics data ($\underline{\mathbf{X}}_{IC}$) is improved through variable selection. Since the number of the original ions ($4587\ variables \times 6\ time\ instants$) compared to the number of cell lines (96) is very large, variable selection needs to be performed to reduce the number of ions by retaining only the ones that maximize the performance of the models.

Variable selection is applied to each model using a backward iterative elimination, which has been used in previous works on this dataset (Barberi, 2023) . In this procedure, three indices are used for variables removal:

- selectivity ratio;
- the beta-value, namely the value of the coefficient for the regressor in the PLS procedure;
- VIP index.

The variables selection is performed as follows:

1. initially the model is cross validated with all available variables;
2. using the indices of variable importance calculated during cross-validation (step 1) three different datasets are created by removing a defined percentage $p = 25\%$ of less important variable;
3. three PLS models are built on the datasets generated at step 2 and cross-validated.
4. only the model (and the associated variables) that demonstrates the best performances (minimum RMSECV) is retained for further iterations. If the exclusion of unimportant variables does not improve the performance of the model (increase in RMSECV) the percentage of variables to remove is decreased and the excluded variables are reinserted into the dataset. The procedure continue with a new iteration (step 1).
5. the procedure is stopped when the exclusion of a single unimportant variable does not produce model performance improvement.

After the variable selection, some additional operations (Table 4.2) are performed to improve the quality of each resulting PLS models. These operations include:

6. excluding a subset of batches during the model building process. Since some parameters are two or more orders of magnitude different with respect to the average parameter value, they are excluded from the PLS models to limit the leverage of these parameters with unusual value. This enhances the robustness of the PLS model and improves the overall quality of the outcomes;
7. logarithmic transformation of the response. The PLS models for factors: Y_{latglc} and Y_{glux} are built on the logarithm of the parameter rather than on the parameter itself. In fact, these parameters exhibit a very large variability that ranges from very low value to very high value. This numerical transformation is able to make the response more normally distributed and allows obtaining better final results.

Table 4.2. *Specific parameter's operations performed before the application of the PLS model.*

parameter	ratio of excluded batches over the total number of batches	$\log_{10}(Y)$
μ_{max}	0	
Y_{mAbglc}	5/96	
Y_{xglc}	5/ 96	
Y_{xglu}	5/96	
KI_{amm}	5/96	
Y_{latglc}	5/96	•
Y_{glux}	5/96	•

The optimized models are able to improve the estimation performance using a lower number of latent variables.

4.2 Validation of the influence of cell line metabolism on cell chemical, physical and biological phenomena

In this Section, the validation performance of the improve PLS models predicting first principle model parameters are presented. Model validation is an essential step in evaluating the effectiveness of PLS models. During this validation process, the performance of a model calibrated using a reduced number of cell lines is evaluated by predicting the behavior of the cell lines that were excluded during calibration. Model validation is conducted through a Monte Carlo procedure, in which 5 cell lines are randomly selected as validation dataset, while the remaining 91 cell lines are used to calibrate the improved model, while the validation dataset is used evaluate the prediction performance of the model. This procedure is repeated for $1 \cdot 10^5$ iterations.

Prediction performance of the PLS models built for each first principle model parameter are reported in Table 4.3.

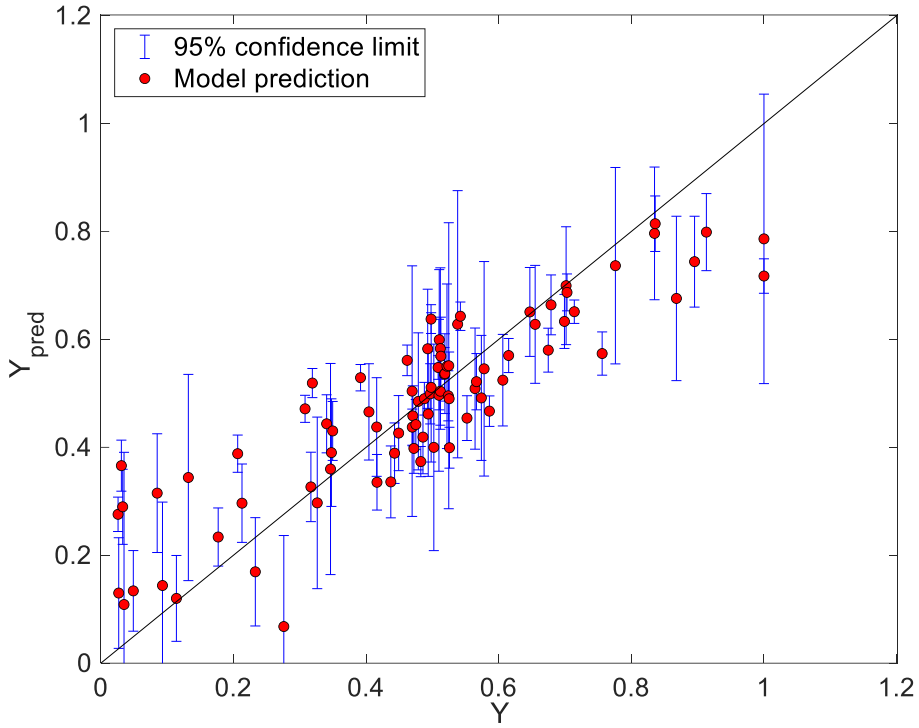
Overall, the estimation performance are satisfactory, as indicated by the high Q^2 index and the Mean Absolute Error (MEA) values that are often much lower than the standard deviation of the parameters values. However, two parameters, namely Y_{mAbglc} and Y_{xglc}, show worse validation performance, resulting in a low Q^2 ($< 50\%$). Lower performance may be due to a reduced correlation between metabolites and Y_{mAbglc}, leading to a small portion of metabolite variability associated to that parameter.

Table 4.3. *PLS model validation results for all first principle model parameters.*

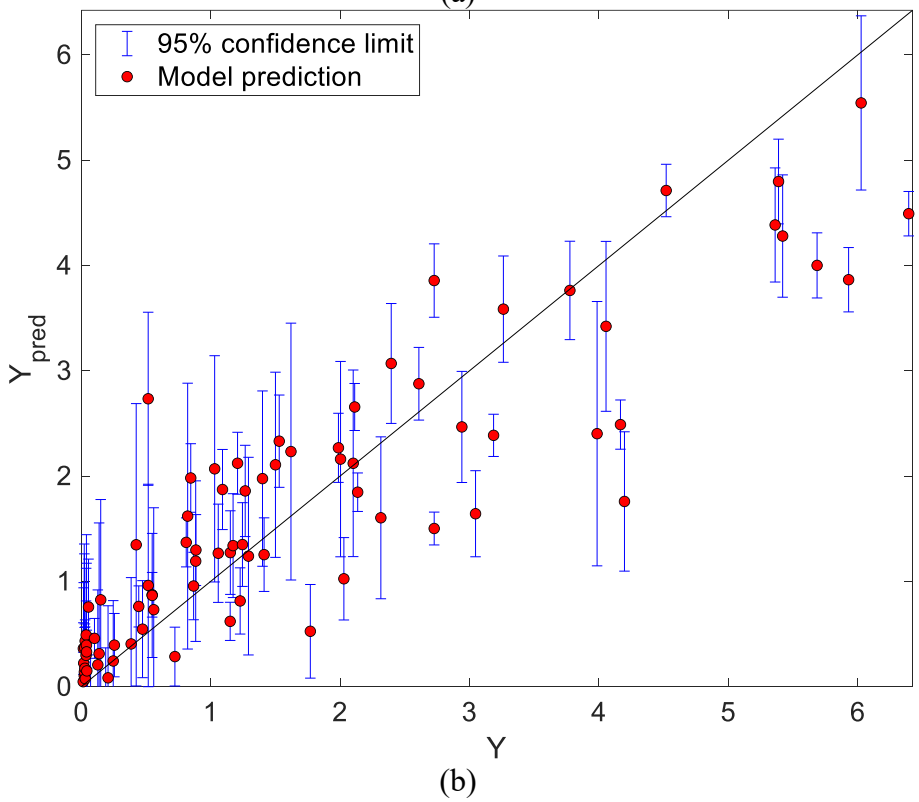
Parameter	μ	σ	Mean of MEA/σ	Median of MEA/σ	Q^2
μ_{max}	0.489	0.255	10.90%	8.64%	75.20%
Y_{mAbglc}	10.500	8.630	8.96%	7.10%	50.00%
Y_{xglc}	1.840	0.710	8.79%	6.67%	45.80%
Y_{xglu}	382.000	348.000	23.30%	16.30%	90.40%
KI_{amm}	1.810	2.290	51.20%	34.70%	76.40%
Y_{latglc}	1.290	1.150	100.00%	41.30%	80.90%
Y_{glux}	3.910	4.850	51.20%	32.90%	82.20%

PLS models results can be further assessed by examining the parity plots in Figure 4.2. The red point represent the average prediction between several iterations, while the blue bars width represent twice the predictions standard deviations. Particularly, Figure 4.2a describe the results for μ_{max}, which exhibits a Q^2 index of 75.2%. The mean predicted values are in good agreement with true parameter values, as indicated by the closeness of the points to the diagonal line. Additionally, the error bands consistently intersects with the plot diagonal, indicating that the error between the true parameter value μ_{max} and the predicted oner is statistically not different from zero. However, it is important to note that higher variability of predictions tend to occur when the parameter exhibit extreme value (close to zero or close to one). This issue could be attributed to the fact that model has few examples at the extreme of the interval; for this reason it tends to extrapolate more parameters value, committing larger error.

In Figure 4.2b, the parity plot represents the results for parameter KI_{amm}. Similarly, to previous case, the Q^2 index is high. However, upon closer examination of the parity plot, it is evident that the first-principle model parameter are not normally distributed. In fact their values are highly concentrated around the zero. For that reason, the PLS models struggles to provide accurate estimations of the parameter when its values are very low. Proper parameter scaling may be useful to improve predictive capability and prediction of value close to zero.


95% confidence limit
Model prediction
Y pred
Y
0
0.2
0.4
0.6
0.8
1
1.2

(a)


95% confidence limit
Model prediction
Y pred
Y
0
1
2
3
4
5
6

(b)

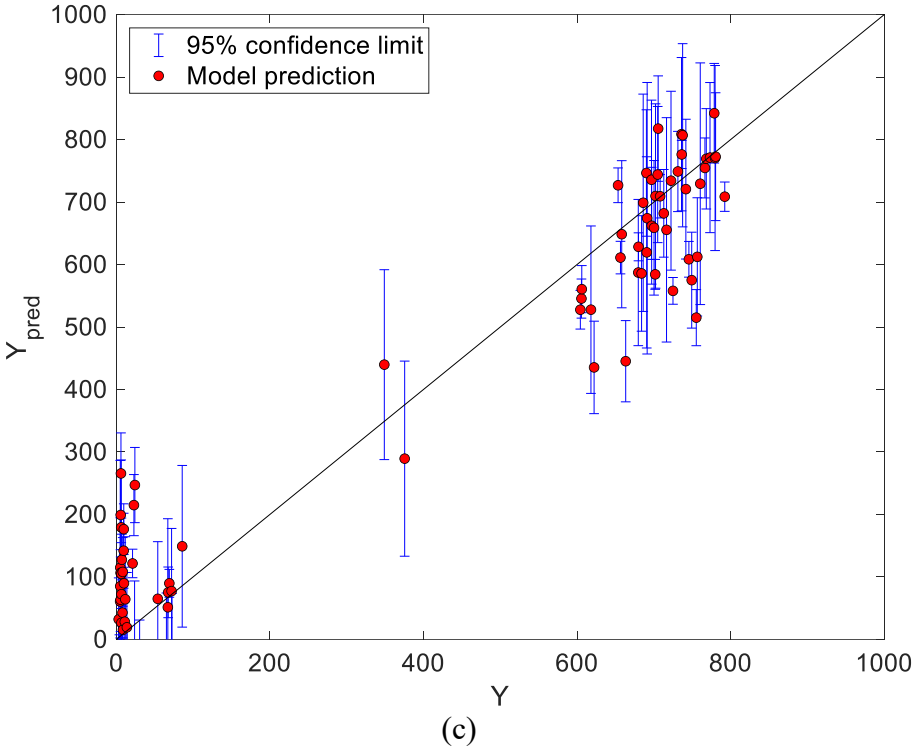


Figure 4.2. *Parity plot PLS validation. (a) μ_{max}, (b) KI_{amm} and (c) Y_{xglu}.*

Additionally, interesting result comes from the observation of Y_{xglu} (Figure 4.2c) In the parity plot, a clear distinction between two clusters appears. The parameter follow a bi-modal distribution. This subdivision is related to the fact that the parameter is extremely sensitive to Glutamate consumption. All high performing cell cultures have a more efficient nutrient consumption and consequently, high value of Y_{xglu}, while low performing cell lines have a less efficient Glutamate consumption. Despite the parameter distribution is not normal, model performances are good ($Q^2 = 90\%$). For this reason, the results of the PLS model for Y_{xglu} can also be used to discriminate cell lines according to their performances.
Parity plots for all the remaining parameters can be found in Appendix C.

4.3 Biological understanding on how CHO cell metabolism is related to cell culture chemical-physical and biological phenomena

In this Section, the results of the PLS model are used to better understand cell metabolism and how it is related to the chemical, physical and biological phenomena occurring into the cell cultures. The data-driven PLS model links metabolites (and thus metabolic traits) to parameters that serve as indicators of a specific biological phenomenon, allowing the investigation of metabolites associated with desired cell culture behaviors. For example, the prediction of μ_{max} from the metabolomics data associates the phenomenon of the cell growth with the metabolism of each cell line through the respective -omics data. This analysis is performed by analyzing the VIP index of the PLS models, namely the ion importance in the prediction of a first-principle

model parameter. In addition, the VIP indices are coupled with the regression parameters (β) to assess how a particular metabolite directly affects the biological phenomena and cell behavior over time.

4.3.1 Biological understanding for μ_{max} – cell growth

In this Section, the metabolites associated with the parameter μ_{max}, which describe cell growth, are presented, and analyzed. The heatmap of Figure 4.3 shows the importance over time for metabolites of μ_{max}.

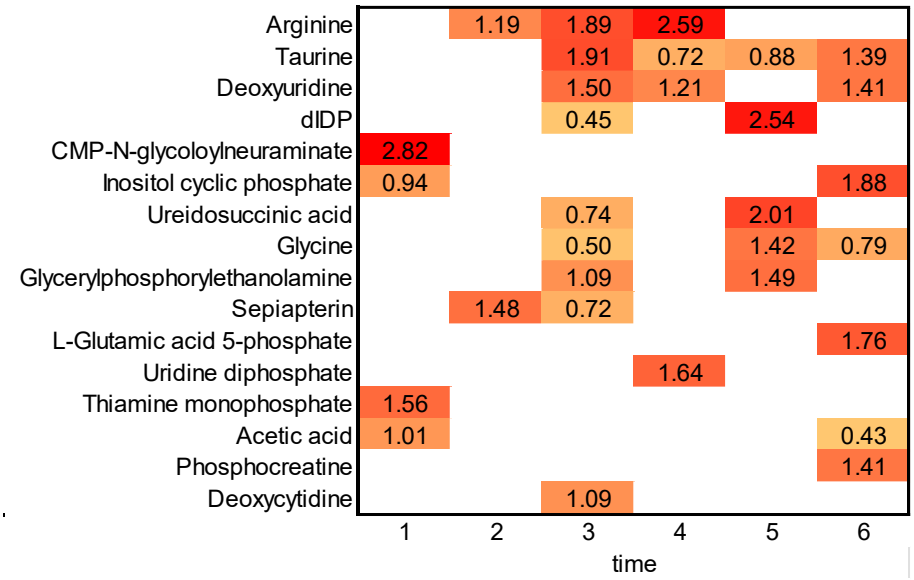


Figure 4.3. *VIP along culture time of metabolites from PLS model predicting μ_{max}. Red color is associated with important metabolites, while a more orange one indicates a less important metabolite.*

Arginine is identified as one of the most important metabolites. The analysis shows that Arginine VIP values are high during the cell growing phase with positive β-values at time 2 and 3. This indicates that high Arginine content during growing phase are correlated to a fast cell population increase. On the other hand, the β is negative at time 4. In fact, high Arginine content in later stages means that the component has not been efficiently consumed; consequently, slower cell grow and lower first-principle model parameter μ_{max} are obtained. This result are consistent to literature findings. Arginine is found to be one of the most significant metabolites associated with the growth of the cell culture (González-Leal et al., 2011).

Also Taurine's VIP values are high, especially at time 3 and 6. Taurine plays a crucial role in the central and final phases of the cell culture (peak of VCC and product formation) with always positive β. This means that higher content of Taurine is associated to large cell growth and consequently high VCC value, and to a more consistent product formation. The biological role of Taurine is in accordance with expectation, since a positive effect on increasing both VCC and Product Titer has been previously observed (M. Liu et al., 2018). Additionally, Taurine has

a relevant impact on all first-principle parameters relating biomass growth to nutrient consumption (i.e., Glucose and Glutamate). These results prove the beneficial effect of Taurine in different cell culture phenomena.
Similar considerations can also be done for Deoxyuridine. High VIP values at times 3, 4 and 6 and positive β of the PLS model in the late part of the cell culture correlate to high component concertation with a fast cell growth. Similarly to Taurine, Deoxyuridine has also been proven to be a positive effect for efficient antibody production (Takagi et al., 2017).
Other metabolites, such as Inositol, Glutamic acid, Uridine and Thiamine phosphate, associated with Glucose metabolism are identified as important in different time instant having a high VIP index. This is something expected, because being Glucose the main nutrient of the cell culture, the byproducts of its metabolism can be associated with the progress of the cell growth.
Finally, dIDP (Diisodecyl phthalate) has a very high VIP index at time 5, and a negative value of the β, meaning that high dIDP concertation correlates with a limited cell growth. This result is consistent to the fact that this metabolite has proven to bring a negative effect on cell cultures growth (Phillips et al., 1982).

4.3.2 Biological understanding for Y_{mAbglc} – antibody production due to Glucose consumption

In this Section, the metabolites associated with the parameter Y_{mAbglc} are presented and discussed. This parameter describes the production of antibody associated with Glucose consumption. The importance over time of metabolites for Y_{mAbglc} is shown in Figure 4.4.
Thiamine monophosphate is found to be a very important metabolite as its VIP values are significantly larger than one from time 1 to 4. Its β values are always positive, indicating that Thiamine is positively correlate to high product formation. Accordingly, the presence of this compound during the first half of the cell culture is identified as having a large impact on the formation of monoclonal antibody.
Propynol adenylate is an essential metabolite for the system under study as indicated by its VIP indices, which are > 1 in the first half of cell culture. The positive βvalues indicate that high Propynol concentration correlated to large product formation and large Y_{mAbglc}. Additionally, Propynol is found to be a key metabolite for parameters Y_{latglc} and Y_{glux} that, similarly to Y_{mAbglc}, regulate the production of other cell components. These observations are consistent with previous studies, which identified Propynol adenylate as an important enzyme involved in the production of essential and complex molecules such as monoclonal antibody, Lactate and Glutamate (D'Ambrosio & Derbyshire, 2020).
Other metabolites, such as 3-Dehydro L-Gulonate, Glyceric acid and Gluconic acid, have a VIP index with increasing importance in late stages of the cell culture. These metabolites, which are

associated to Glucose metabolism, highlight the important relationship between the Glucose and antibody formation.

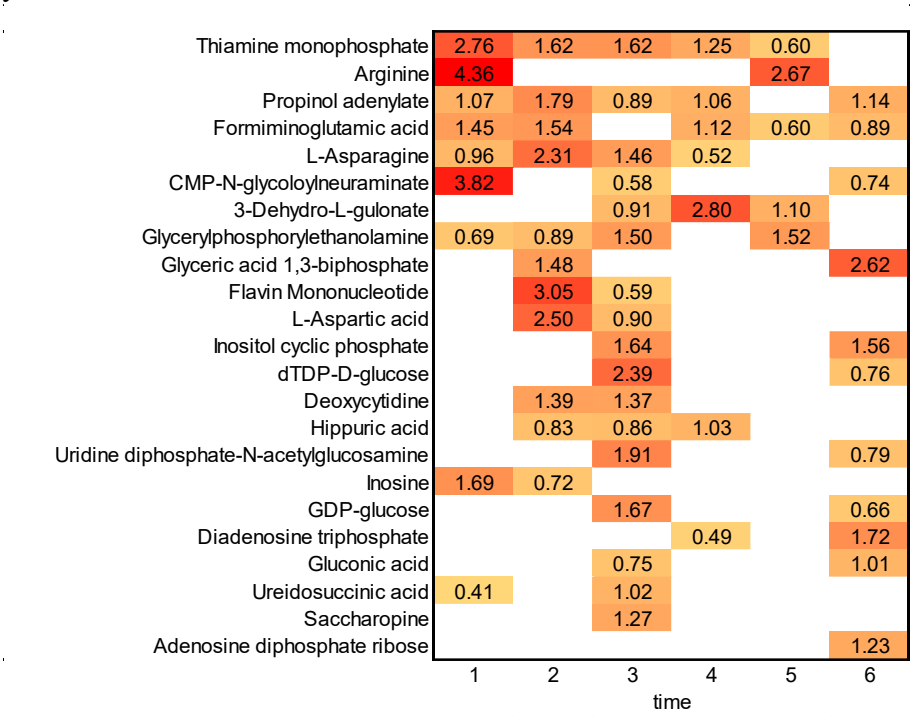


Figure 4.4. *VIP along culture time of metabolites from PLS model predicting* Y_{mAbglc}. *Red color is associated with important metabolites, while a more orange one indicates a less important metabolite.*

Additionally, Formiminoglutamic acid has high VIP values during the exponential growth and stationary phase. The accumulation of this compound has been identified as mainly related to Glutamate consumption in previous studies (Coulet et al., 2022). This indicates that there is also a relationship between Glutamate and product formation, even though antibody production is modeled only as a result of glucose consumption. This fact may be due to the simplified nature of the monoclonal antibody model balance, which makes the parameter also capture lumped metabolic traits associated with antibody production. This suggests a refinement of the model by improving the product formation associated with Glutamate consumption.

4.3.3 Biological understanding for KI_{amm} *– Ammonia inhibition in the cell culture*

In this Section, the metabolites associated with the parameter KI_{amm}, which describe Ammonia inhibition, are identified. The heatmap of Figure 4.5 shows the importance over time of metabolites for KI_{amm} prediction.

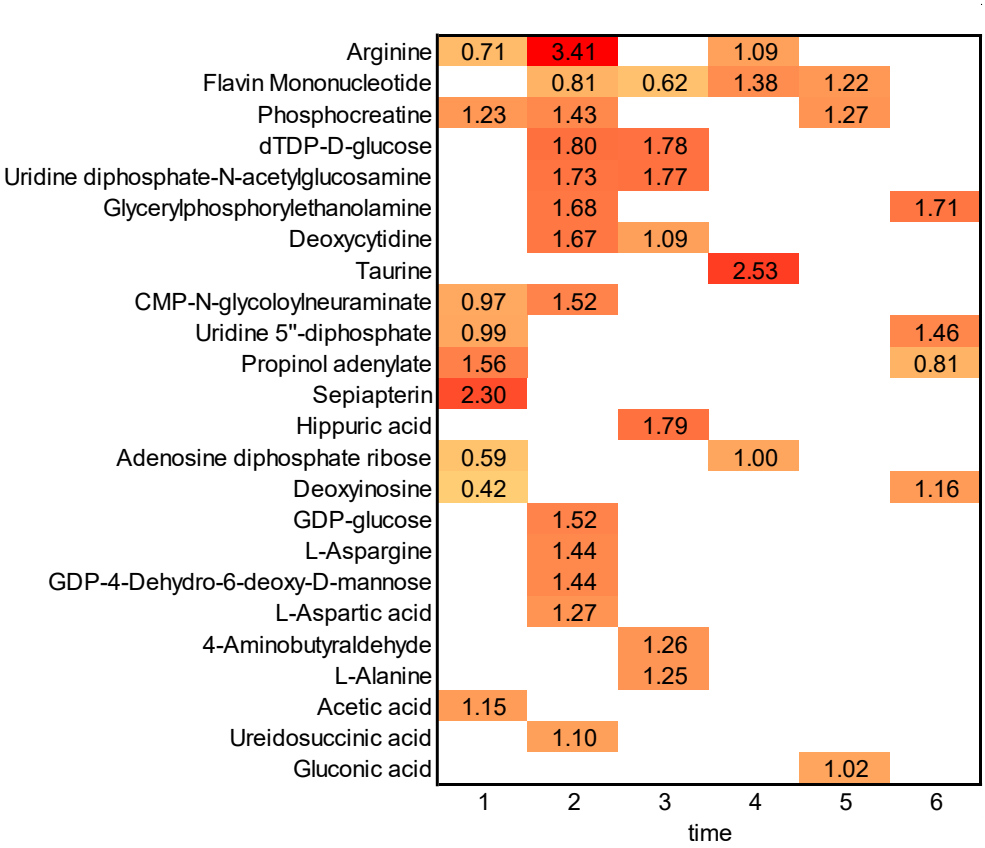


Figure 4.5. *VIP along culture time of metabolites from PLS model predicting* KI_{amm}*. Red color is associated with important metabolites, while a more orange one indicates a less important metabolite.*

Arginine is found to be one of the most important metabolites associate to KI_{amm}, having a high VIP at time 2 and 3. The β values associated to the metabolite are positive, indicating that low concentration of Arginine is correlated to low values of the parameter and consequently to a larger Ammonia inhibition. In the studied system, low Arginine levels resulted in a reduction in cell resistance to Ammonia and consequently slow culture growth at high Ammonia concentrations. This result is consistent with previous studies, which showed that Arginine deprivation can cause the death of the cell culture (Scott & al, 2000).

Different amine (i.e., Glucosamine and Ethanolamine) and aminate compounds have large VIP values in the first half of the cell culture ($VIP > 1$). Their β are positive, indicating that high concentration of these metabolites are associated to large parameter values and consequently to a high resistance of the system towards Ammonia, suggesting that having greater amount of Nitrogen stored other compounds rather than as Ammonia, reduced the inhibition of cell growth due to ammonia.

4.3.4 Biological understanding for Y_{xglc} – biomass growth due to Glucose consumption

This Section presents and analyzes the metabolites associated with Y_{xglc}, which describes the yield of biomass associate to Glucose consumption. The heatmap of Figure 4.6 shows the importance over time of metabolites for Y_{xglc}.

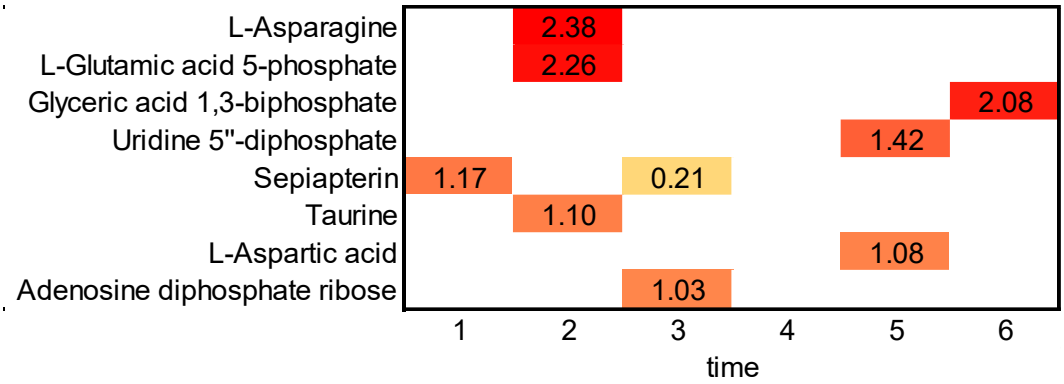


Figure 4.6. *VIP along culture time of metabolites from PLS model predicting Y_{xglc} Red color is associated with important metabolites, while a more orange one indicates a less important metabolite.*

In this case, all identified metabolites have high VIP values in single time points. L-Glutamic acid has large VIP index at time 2, ADP at time 3, Uridine-diphosphate at time 5, and Glyceric acids at time 6. This phenomenon may be related to the fact that many intermediates are involved in the metabolism of Glucose, and different intermediates can be influential at different instants. Monitoring these compounds provides the capability to assess cell performance with respect to Glucose consumption. Additionally, L-Asparagine has very high VIP value at time 2, highlighting it is a good predictor of Glucose consumption. Furthermore, its negative β value, indicates that L-Asparagine content is anticorrelated to Glucose consumption, meaning that the higher L-Asparagine concentrations, the lower the parameter values, and the lower the Glucose consumption. Furthermore, Asparagine VIP is not negligible for the parameter that controls the production of antibody, Y_{mAbglc},. The results are consistent with what has been reported in previous studies, where Asparagine, being a standard amino acid used in protein synthesis, provided useful indication on the production of monoclonal antibody (Duarte et al., 2014).

4.3.5 Biological understanding for Y_{xglu} – biomass growth due to Glutamate consumption

This Section presents and analyzes the metabolites associated with Y_{xglu}, which describes the yield of biomass associated with Glutamate consumption. The heatmap of Figure 4.7 shows the importance over time of metabolites for Y_{xglu} prediction.

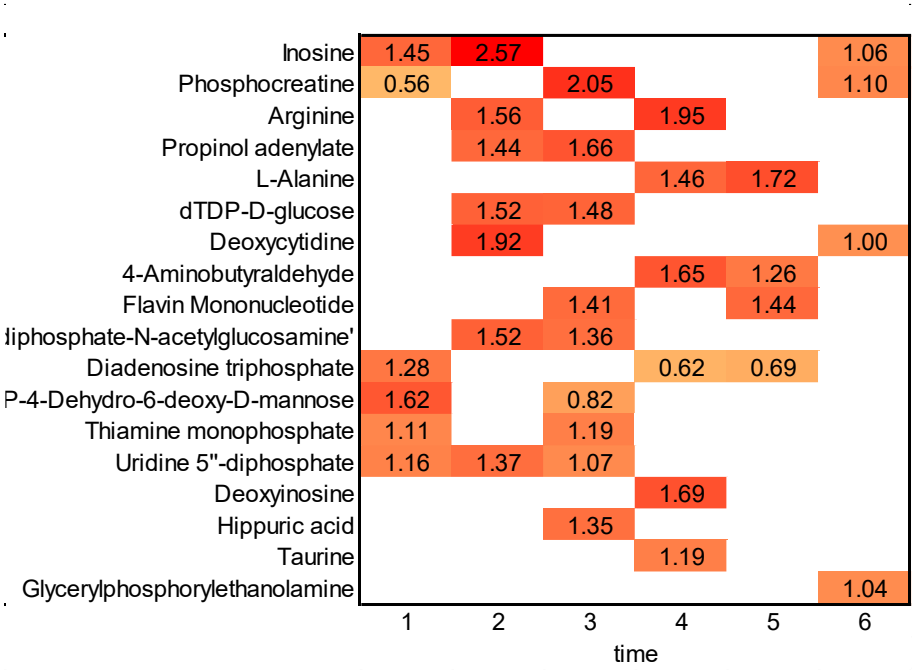


Figure 4.7. *VIP along culture time of metabolites from PLS model predicting Y_{xglu}. Red color is associated with important metabolites, while a more orange one indicates a less important metabolite.*

Inosine is a very important metabolite in the initial part of the cell culture, when large Glutamate consumption is occurring. Inosine β values have negative values meaning that high levels of this metabolite are associated with low Glutamate consumption. Consequently, high performing cell cultures (characterized by larger Y_{xglu} values), show low levels of Inosine. Inosine significance is consistent to literature expectation, since it was found to be an alternative carbon source when limited Glucose is available (Wang et al., 2020). Accordingly, the high importance of Inosine associated with Glutamate consumption confirms the key role of Glutamate as nutrient for cell cultures.

Other compounds, related to cell energy storage and viability, are important for Y_{xglu}. Of those, Phosphocreatine, dTDP-D-glucose, and others phosphate components have non negligible VIP values. This further supports the observation that Glutamate plays a crucial role as backup nutrient for cell cultures.

Additionally, the presence of metabolites such as Propinol Adenylate and Arginine suggest a relationship between Glutamate and the synthesis of monoclonal antibodies, highlighting the dual nature of Glutamate's role in cell culture.

4.3.6 Biological understanding for Y_{latglc} – Lactate production due to Glucose consumption

This Section reported the metabolites associated with the parameter Y_{latglc}, which describe the yield of Lactate due to Glucose consumption. The importance over time of metabolites for Y_{latglc} prediction is shown in Figure 4.8.

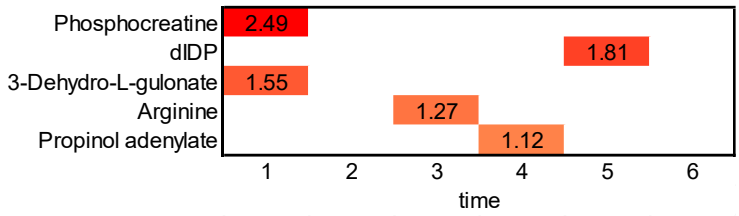


Figure 4.8. *VIP along culture time of metabolites from PLS model predicting Y_{latglc} Red color is associated with important metabolites, while a more orange one indicates a less important metabolite.*

Among these, Phosphocreatine and Gulonate, which are Glucose derivatives, have negative β values. This means that large amount of these metabolites correlates to a low value of the parameter and to low Lactate production due to the consumption of Glucose. Having multiple Glucose intermediates as important indicators of Lactate behavior suggests that the course of Glucose metabolism is an important indicator of Lactate production.

Diisodecyl phthalate (dIDP) has positive β, indicating that high metabolite levels are linked to a larger Lactate secretion, which limits the growth of the culture and increases cell death, as supported by previous studies, which identified that dIDP is a component usually associated with cell death (Section 4.3.1).

4.3.7 Biological understanding for Y_{glux} – Glutamate production due to cell activity

In this Section, the metabolites associated with the parameter Y_{glux}, which describe Glutamate production due to cell activity, are presented and analyzed. The heatmap of Figure 4.9 shows the importance over time of metabolites for Y_{glux} prediction.

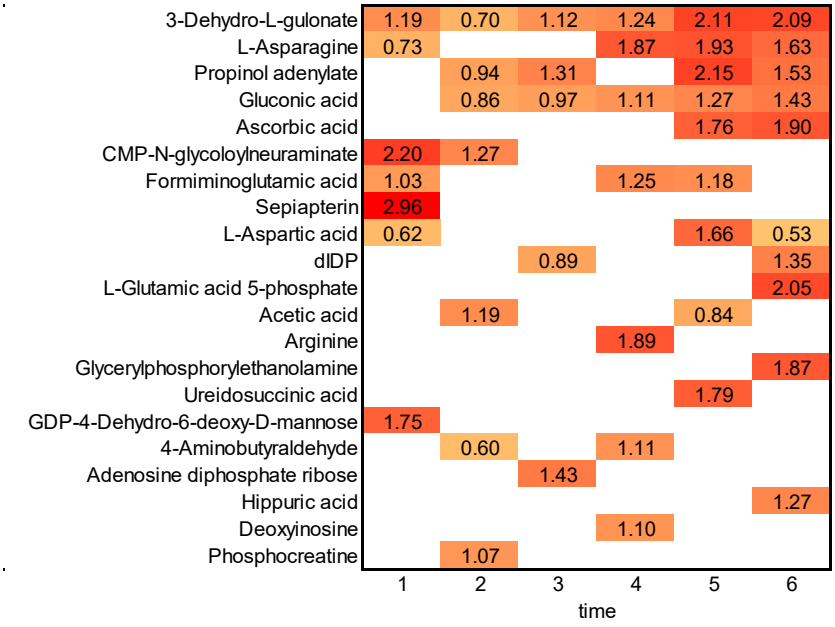


Figure 4.9. *VIP along culture time of metabolites from PLS model predicting Y_{glux} Red color is associated with important metabolites, while a more orange one indicates a less important metabolite.*

VIP values for L-Gulonate are high in all time instants, suggesting that it is an important indicator of the metabolic pathway associated with Glutamate production. Literature studies show that L-Gulonate is a component produced by Glucose metabolism (Grindley et al., 1988). This may suggest that Glucose intermediates might be somehow involved in Glutamate production during cell activity.

Similarly, Gluconic acid and Ascorbic acid that have high VIP index in the late stages of the cell culture. Their β are always negative suggesting that low metabolites levels are associated to high Glutamate production. Since these components are directly produced from Glucose (Yadav et al., 2022), this enforces the relationships between Glucose intermediates and Glutamate synthesis.

Other important components in determining Y_{glux} are Asparagine and Propynol Adenylate. As said before (Section 4.3.2), these metabolites have relevant VIP values for other parameters, such as Y_{latglc} and Y_{mAbglc}. Accordingly, these metabolites have a central role in determining the synthesis of several cell compounds, such as cell components and proteins like monoclonal antibodies.

Conclusion

The primary objective of this Thesis is to explore an innovative way to integrate metabolomics data into the framework of first-principles models for CHO cell lines. This integration aims to gain valuable insights into processes by examining it from a new perspective and reveal the relationships between chemical, physical and biological system phenomena and cell metabolism.

In this context, a new model for CHO cells is proposed that enhances existing Literature models by refining the role of Glutamate and Lactate. The model includes additional contributions by modifying the governing equations of the system. Structural identifiability results show that all parameters of the model can be estimated from process data.

The proposed model effectively captures the behavior of important process variables, such as Viable Cell concentration and Product titer, and provides improved results for fitting the system under study. The estimation results are obtained by fitting only seven parameters while keeping the fixed parameters constant. By carefully selecting the most significant parameters to be estimated and assigning appropriate values to the fixed parameters, the overall fitting results are quite satisfactory (Chapter 3). At this phase, the metabolomics data are integrated into the procedure and bridged to first-principles model parameters using a PLS model. The PLS model is accurately calibrated by preprocessing the regressor dataset and by retaining only the most informative ions (Chapter 4). The outcomes of this process reveal a robust association between the metabolites and parameter values, which embeds chemical-physical and biological phenomena characteristic of the system.

The main objective of the Thesis can be achieved by examining this relationship, as it provides access to valuable new insights on the system under study. The most relevant regressors (ions) associated with a parameter are identified using the VIP value that allows understanding of which metabolites and metabolic traits are most relevant to a particular phenomenon. Additionally, a cross-consideration of the PLS beta values allows understanding of the effect of each metabolite on the parameter value, specifically the positive or negative effect in enhancing or not a particular cellular phenomenon.

Final results are available only for the subset of the 7 most characteristic parameters, since the main limitation of this work is the limited number of experimental points, nevertheless the obtained results can serve as a valuable starting point for new considerations related to the CHO cell system. An additional constraint is mainly related to the type of cellular system used. Enhancing the CHO models can provide significant improvements by strengthening the relationships between the model's parameters and the metabolites.

This approach, however, is considered quite flexible and can be utilized to study one or more cellular phenomena at a time. The potential of this method can easily be improved by having many experimental points. Moreover, this strategy allows to approach the problem in an innovative way by linking metabolomics data to first-principles model parameters. This offers the possibility to investigate the system from a new perspective and generate new insights that can be used to improve this type of system.

Appendix A

Proposed CHO cellular model

The proposed CHO cell model improves the existing model (Kontoravdi et al., 2010b) for describing CHO cell lines. The improvement of the model mainly concerns the role of Lactate and Glutamate, but all the introduced changes are fully described in Section 3.4. In this Section, the complete structure of the model is presented.

The model proposed and used in this work is composed of:

- eight state variables (Table A.1) :$V, C_{glc}, C_{glu}, C_{lat}, C_{amm}, X_V, C_{mAb}, C_{gln}$
- four input conditions (Table A.1): $F_{IN}, F_{OUT}, C_{glc,in}, C_{glu,in}$
- twenty-five parameters (Table A.2Table A.).

Table A.1. *Proposed CHO cell model. List of variables used by the model.*

	Variable
V	Fedbatch volume
C_{glc}	Glucose concentration
C_{glu}	Glutamate concentration
C_{lat}	Lactate concentration
C_{amm}	Ammonia concentration
X_V	Viable cell concentration
C_{mAb}	Product titer
C_{gln}	Glutamine concentration
F_{IN}	Inlet flowrate
F_{OUT}	Outlet flowrate
$C_{glc,in}$	Feed Glucose concentration
$C_{glu,in}$	Feed Glutamate concentration

Table A.2. *Proposed CHO cell model. List of parameters used by the model.*

	Parameter
$K_{c,lat}$	Control factor to Lactate consumption (high Lactate concentration)
$K_{c,glc}$	Control factor to Lactate consumption (low Glucose concentration)
Y_{ammgln}	Yield of Ammonia production due to Glutamine consumption
Y_{ammglu}	Yield of Ammonia production due to Glutamate consumption
Y_{latglc}	Yield of Lactate production due to Glucose consumption
Y_{latglu}	Yield of Lactate production due to Glutamate consumption
Y_{mAbglc}	Yield of Product formation due to Glucose consumption
Y_{xlat}	Yield of biomass growth due to Lactate consumption
k_1	Glutamate to Glutamine constant
k_2	Glutamine to Glutamate constant
K_{glc}	Glucose contribution to cell growth
K_{glu}	Glutamate contribution to cell growth
KI_{lat}	Lactate contribution to cell inhibition
KI_{amm}	Ammonia contribution to cell inhibition
$K_{d,lat}$	Lactate contribution to cell death
$K_{d,amm}$	Ammonia contribution to cell death
μ_{max}	Maximum cell growth rate
$\mu_{d,max}$	Maximum cell death rate
Y_{xglc}	Yield of biomass growth due to Glucose consumption
m_{glc}	Glucose maintenance factor
Y_{xglu}	Yield of biomass growth due to Glutamate consumption
m_{glu}	Glutamate maintenance factor
Y_{xgln}	Yield of biomass growth due to Glutamine consumption
m_{gln}	Glutamine maintenance factor
Y_{glux}	Yield of Glutamate production due to cell activity

The complete set of equation is here reported:

$$\frac{dV}{dt} = F_{IN} - F_{OUT} \quad (4.1)$$

$$\frac{dX_V}{dt} = (\mu - \mu_D)X_V - \frac{F_{IN}}{V}X_V \quad (4.2)$$

$$\mu = \mu_{max}\left(\frac{C_{GLC}}{K_{GLC} + C_{GLC}}\right)\left(\frac{C_{GLU}}{C_{GLU} + K_{GLU}}\right)\left(\frac{KI_{lat}}{KI_{lat} + C_{lat}}\right)\left(\frac{KI_{amm}}{KI_{amm} + C_{amm}}\right) \quad (4.2a)$$

$$\mu_D = \mu_{D,max}\left(\frac{C_{amm}^{\alpha_n}}{C_{amm}^{\alpha_n} + K_{D,amm}^{\alpha_n}}\right)\left(\frac{C_{lat}}{C_{lat} + K_{D,amm}}\right) \quad (4.2b)$$

$$\frac{dC_{GLC}}{dt} = \frac{F_{IN}}{V}\left(C_{GLC,IN} - C_{GLC}\right) - Q_{GLC}X_V \quad (4.3)$$

$$Q_{GLC} = \frac{\mu}{Y_{x,GLC}} + m_{GLC} \quad (4.3a)$$

$$\frac{dC_{LAT}}{dt} = -\frac{F_{IN}}{V}C_{LAT} + Q_{lat,glc}X_V + Q_{lat,glu}X_V - Q_{lat,cons}X_V \quad (4.4)$$

$$Q_{lat,glc} = Q_{GLC}Y_{lat,glc} \quad (4.4a)$$

$$Q_{lat,glu} = Q_{GLU} Y_{lat,glu} \quad (4.4b)$$

$$Q_{lat,cons} = \frac{1}{Y_{x,lat}} \left(\frac{C_{lat}}{K_{c,lat} + C_{lat}} \right) \left(\frac{K_{c,glu}}{K_{c,glu} + C_{GLU}} \right) \quad (4.4c)$$

$$\frac{dC_{GLU}}{dt} = \frac{F_{IN}}{V} \left(C_{in,glu} - C_{GLU} \right) - Q_{GLU} X_V + Q_{glu,x} + k_1 C_{GLN} - k_2 C_{GLU} C_{AMM} \quad (4.5)$$

$$Q_{GLU} = \frac{\mu}{Y_{x,glu}} + m_{GLU} \quad (4.5a)$$

$$Q_{glu,x} = \frac{\mu}{Y_{glu,X}} \quad (4.5b)$$

$$\frac{dC_{GLN}}{dt} = -\frac{F_{IN}}{V} C_{GLN} - k_1 C_{GLN} + k_2 C_{GLU} C_{AMM} - Q_{GLN} X_V \quad (4.6)$$

$$Q_{GLN} = \frac{\mu}{Y_{x,gln}} + m_{gln} \quad (4.6a)$$

$$\frac{dC_{AMM}}{dt} = -\frac{F_{IN}}{V} C_{AMM} + k_1 C_{GLN} - k_2 C_{GLU} C_{AMM} + Q_{amm,glu} X_V + Q_{amm,gln} X_V \quad (4.7)$$

$$Q_{amm,glu} = Y_{amm,glu} Q_{glu} \quad (4.8)$$

$$Q_{amm,gln} = Y_{amm,gln} Q_{gln} \quad (4.9)$$

$$\frac{dC_{mAb}}{dt} = -\frac{F_{OUT}}{V} C_{mAb} + Q_{mAb} X_V^2 \quad (4.30)$$

$$Q_{mAb} = Y_{mAb,glc} Q_{GLC} \quad (4.10a)$$

Appendix B

Results of the EET sensitivity analysis

In this Section, all the results of the EET sensitivity analysis performed in this Thesis are reported.

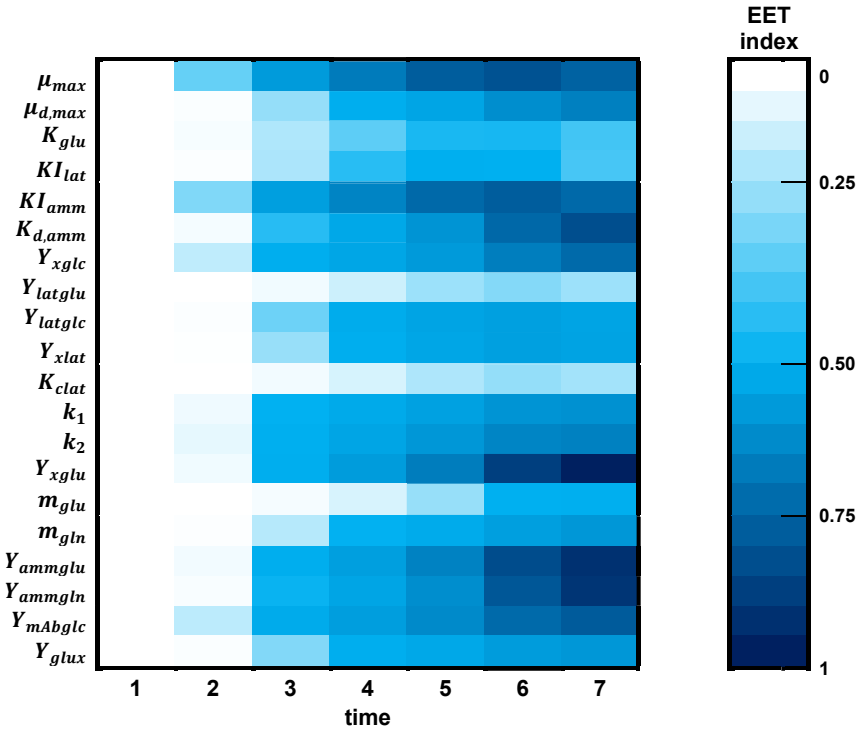


Figure B.1. *Results of the EET sensitivity analysis for the Product Titer.*

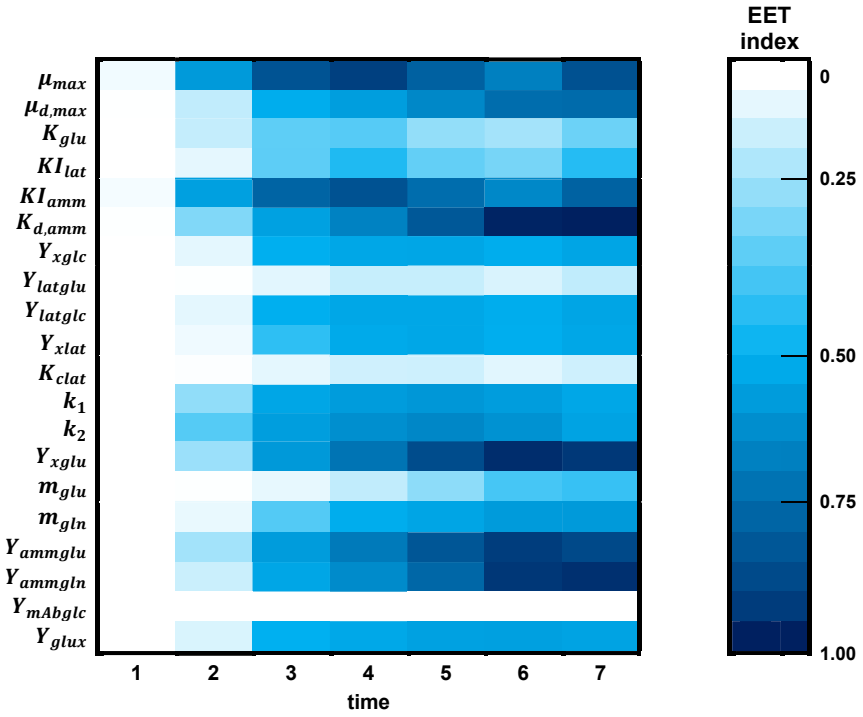


Figure B.2. *Results of the EET sensitivity analysis for the Viable Cell Concentration.*

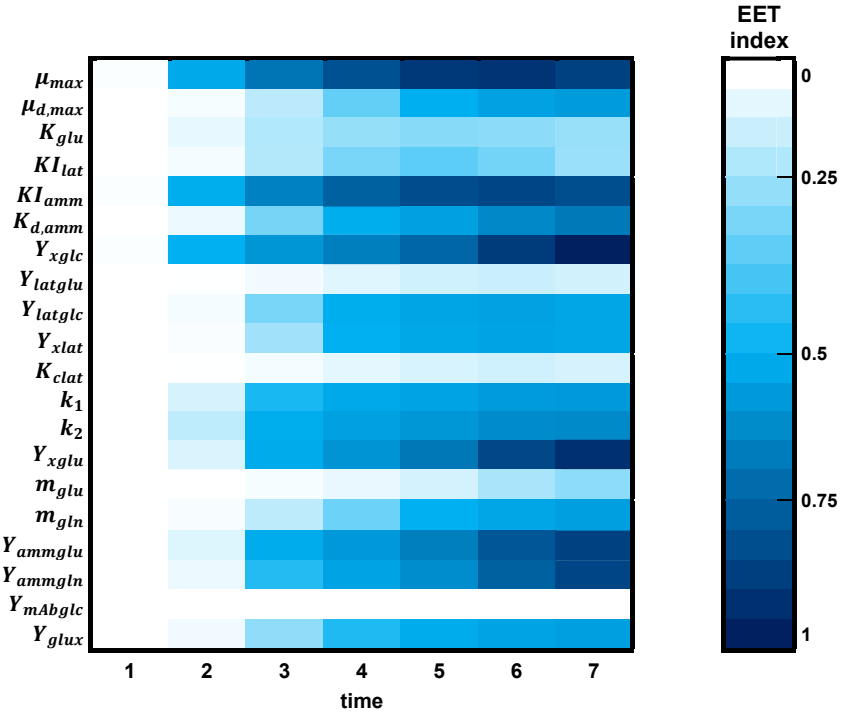


Figure B.3. *Results of the EET sensitivity analysis for the Glucose..*

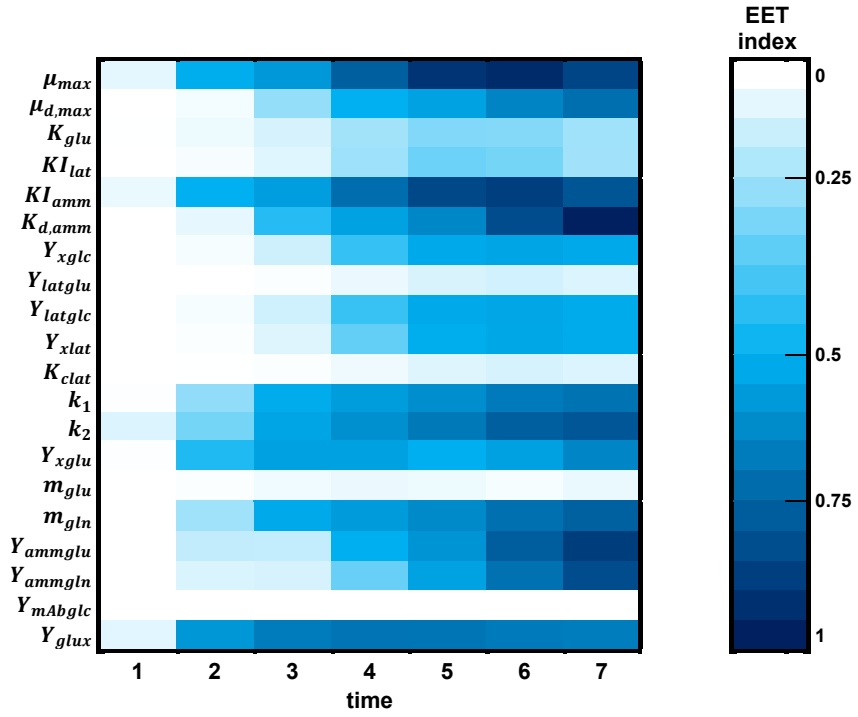


Figure B.4. *Results of the EET sensitivity analysis for the Glutamate.*

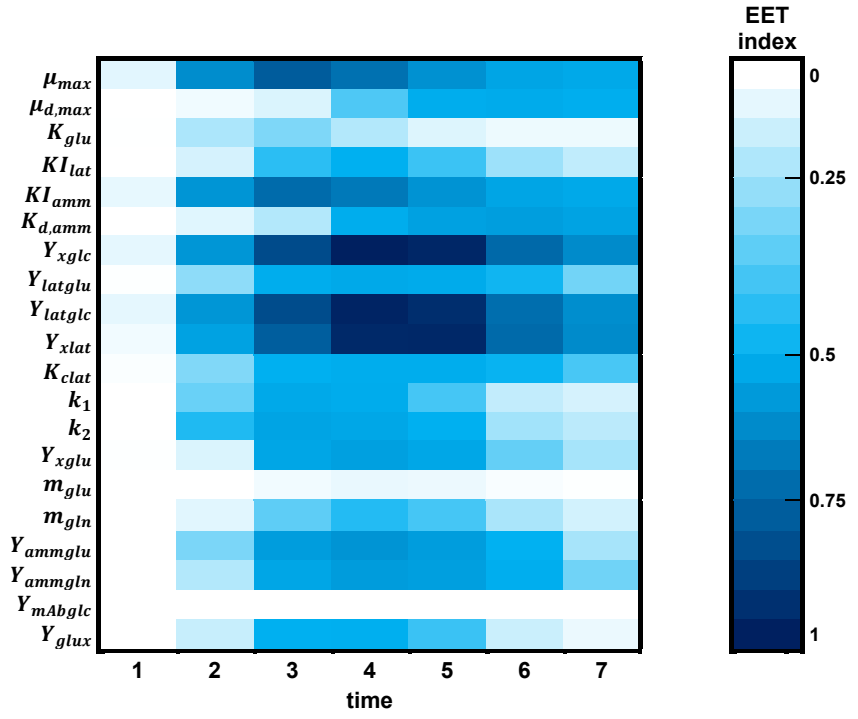


Figure B.5. *Results of the EET sensitivity analysis for the Lactate.*

Appendix C

Parity plot for PLS validation

In this Section, all the parity plots derived for all the PLS models developed in this Thesis are presented.

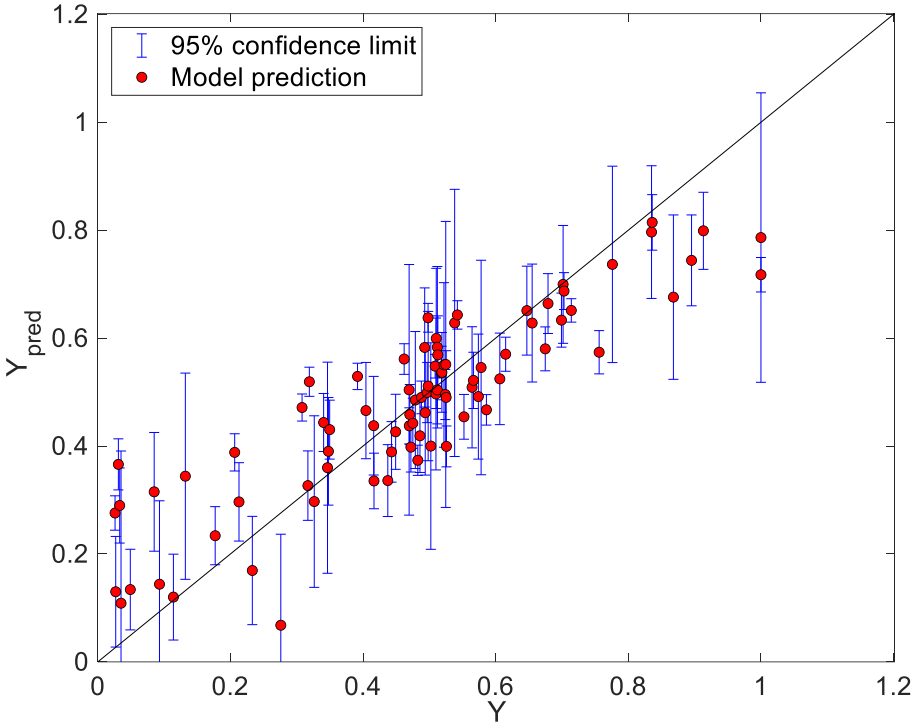


Figure C.1. *Parity plot PLS validation for* μ_{max}

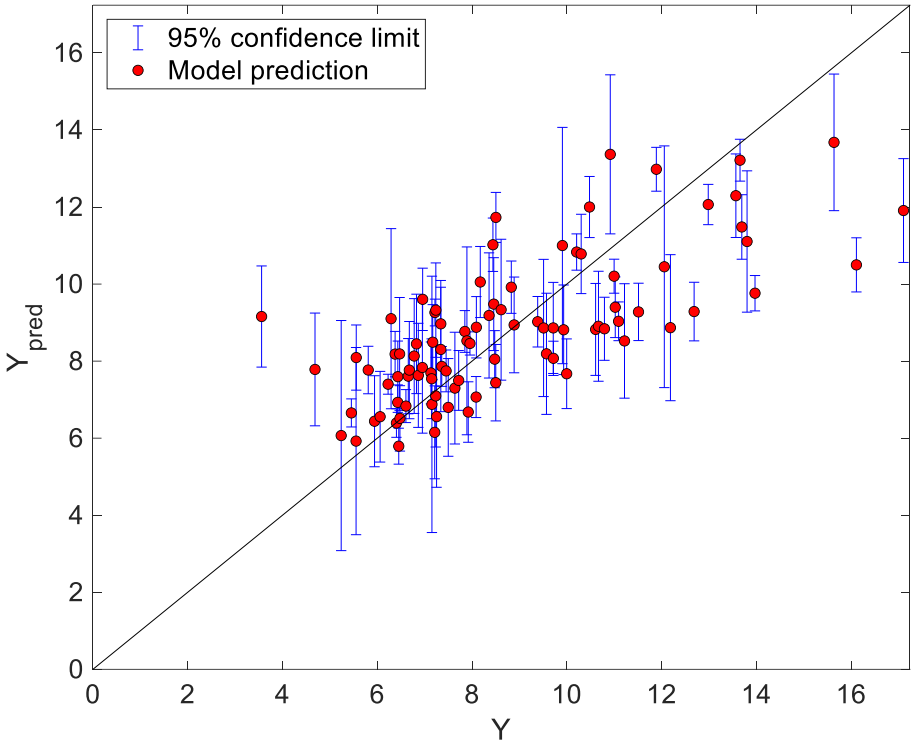


Figure C.2. *Parity plot PLS validation for* Y_{mAbglc}

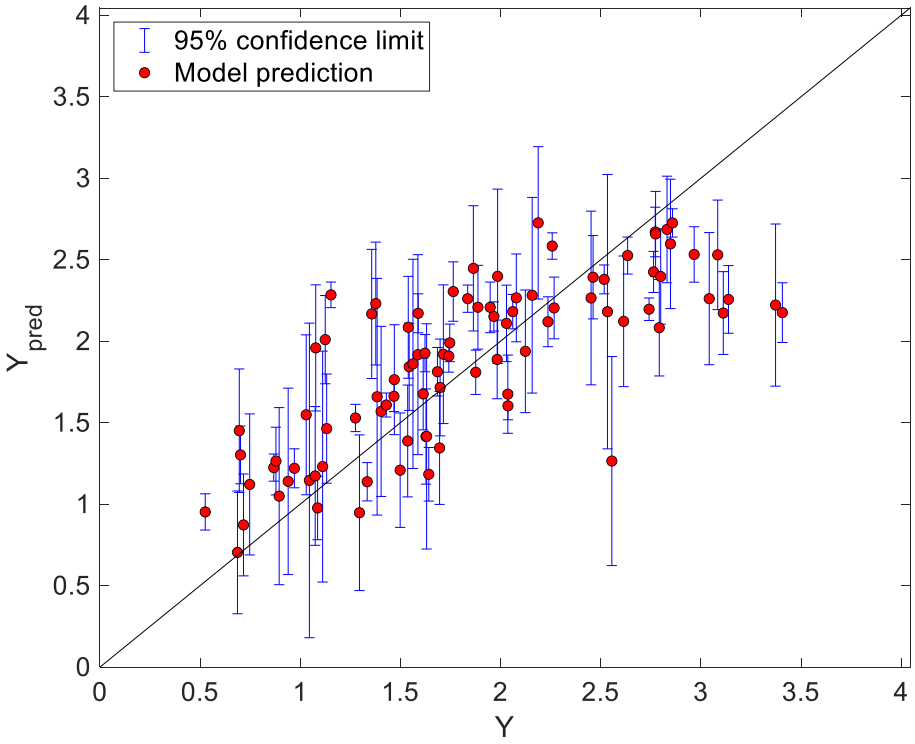


Figure C.3. *Parity plot PLS validation for* Y_{xglc}

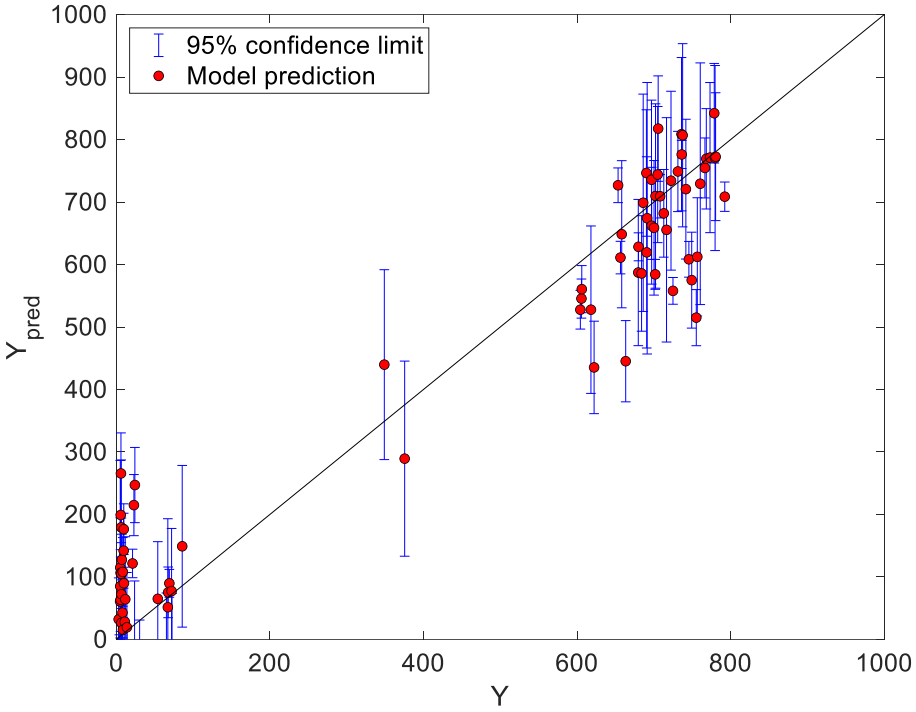


Figure C.4. *Parity plot PLS validation for* Y_{xglu}

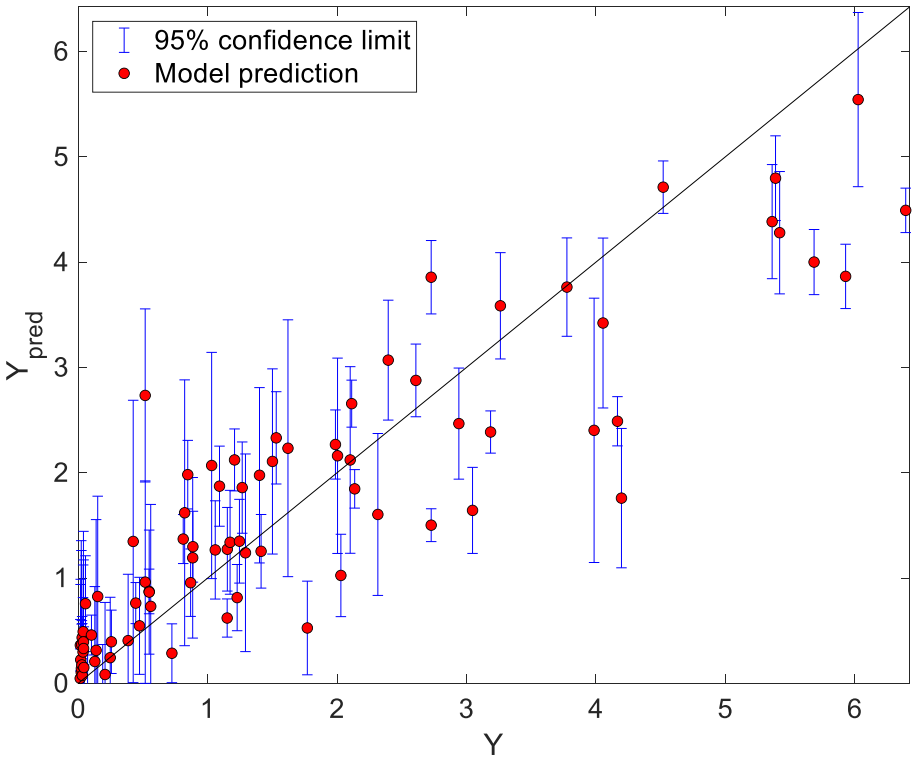


Figure C.5. *Parity plot PLS validation for* KI_{amm}

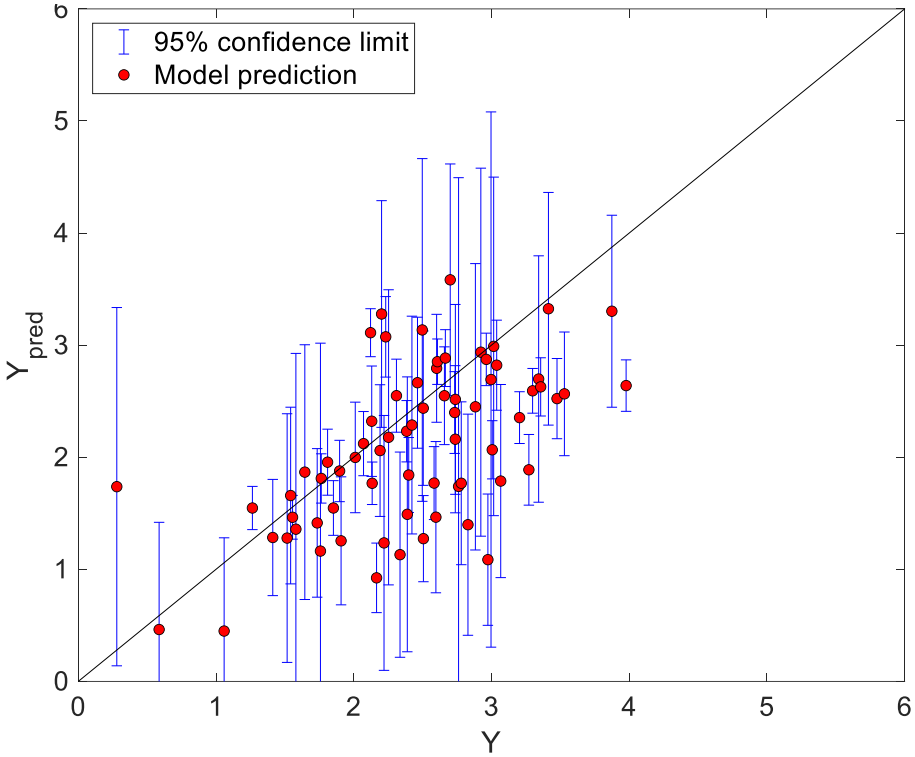


Figure C.6. *Parity plot PLS validation for* Y_{latglc}

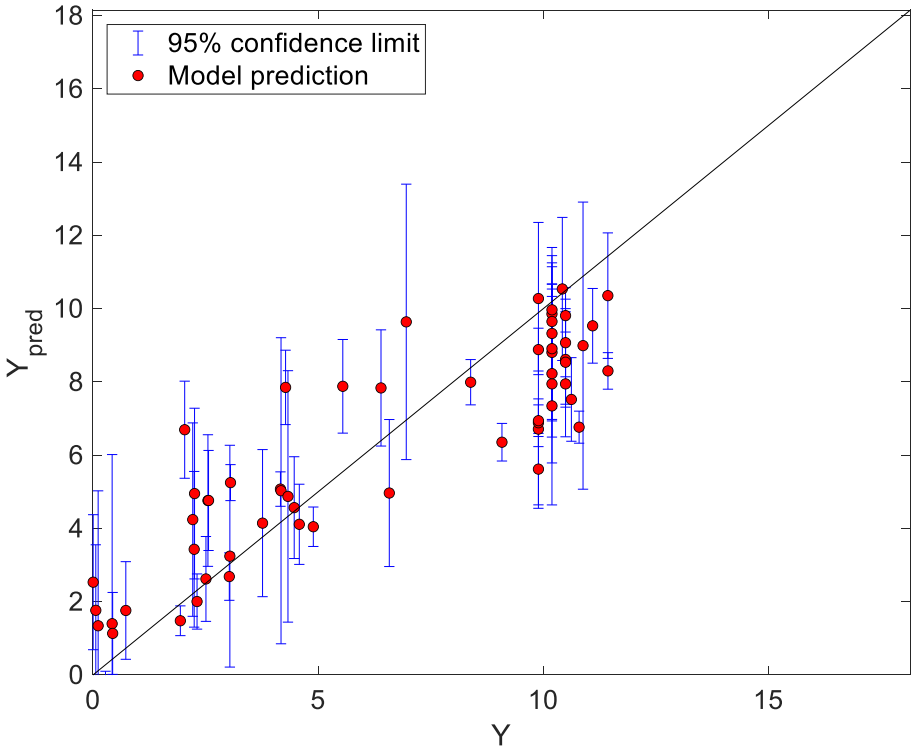


Figure C.7. *Parity plot PLS validation for* Y_{glux}

References

Ahn, W. S., & Antoniewicz, M. R. (2012). Towards dynamic metabolic flux analysis in CHO cell cultures. *Biotechnology Journal*, *7*(1), 61–74. https://doi.org/10.1002/BIOT.201100052

Atangana, A. (2017). Fractional Operators with Constant and Variable Order with Application to Geo-Hydrology. *Fractional Operators with Constant and Variable Order with Application to Geo-Hydrology*, 1–396. https://doi.org/10.1016/C2015-0-05711-2

Barberi, G. (2023). *DIGITAL MODELS TO SUPPORT MONOCLONAL ANTIBODIES DEVELOPMENT IN THE BIOPHARMACEUTICAL INDUSTRY 4.0*.

Barberi, G., Benedetti, A., Diaz-Fernandez, P., Sévin, D. C., Vappiani, J., Finka, G., Bezzo, F., Barolo, M., & Facco, P. (2022a). Integrating metabolome dynamics and process data to guide cell line selection in biopharmaceutical process development. *Metabolic Engineering*, *72*, 353–364. https://doi.org/10.1016/J.YMBEN.2022.03.015

Barberi, G., Benedetti, A., Diaz-Fernandez, P., Sévin, D. C., Vappiani, J., Finka, G., Bezzo, F., Barolo, M., & Facco, P. (2022b). Integrating metabolome dynamics and process data to guide cell line selection in biopharmaceutical process development. *Metabolic Engineering*, *72*, 353–364. https://doi.org/10.1016/J.YMBEN.2022.03.015

Borgonovo, E., Castaings, W., & Tarantola, S. (2011). Moment Independent Importance Measures: New Results and Analytical Test Cases. *Risk Analysis*, *31*(3), 404–428. https://doi.org/10.1111/J.1539-6924.2010.01519.X

Botton, A., Barberi, G., & Facco, P. (2022). Data Augmentation to Support Biopharmaceutical Process Development through Digital Models—A Proof of Concept. *Processes 2022, Vol. 10, Page 1796*, *10*(9), 1796. https://doi.org/10.3390/PR10091796

Campolongo, F., Cariboni, J., & Saltelli, A. (2007). An effective screening design for sensitivity analysis of large models. *Environmental Modelling & Software*, *22*(10), 1509–1518. https://doi.org/10.1016/J.ENVSOFT.2006.10.004

Castelli, M. S., McGonigle, P., & Hornby, P. J. (2019). The pharmacology and therapeutic applications of monoclonal antibodies. *Pharmacology Research & Perspectives*, *7*(6), e00535. https://doi.org/10.1002/PRP2.535

Chiu, M. L., Goulet, D. R., Teplyakov, A., & Gilliland, G. L. (2019). Antibody Structure and Function: The Basis for Engineering Therapeutics. *Antibodies (Basel, Switzerland)*, *8*(4). https://doi.org/10.3390/ANTIB8040055

Chong, I. G., & Jun, C. H. (2005). Performance of some variable selection methods when multicollinearity is present. *Chemometrics and Intelligent Laboratory Systems*, *78*(1–2), 103–112. https://doi.org/10.1016/J.CHEMOLAB.2004.12.011

Coulet, M., Kepp, O., Kroemer, G., & Basmaciogullari, S. (2022). Metabolic Profiling of CHO Cells during the Production of Biotherapeutics. *Cells*, *11*(12). https://doi.org/10.3390/CELLS11121929/S1

D'Ambrosio, H. K., & Derbyshire, E. R. (2020). Investigating the Role of Class i Adenylate-Forming Enzymes in Natural Product Biosynthesis. *ACS Chemical Biology*, *15*(1), 17–27. https://doi.org/10.1021/ACSCHEMBIO.9B00865/ASSET/IMAGES/MEDIUM/CB9B00865_0006.GIF

Dean, J., & Reddy, P. (2013). Metabolic analysis of antibody producing CHO cells in fed-batch production. *Biotechnology and Bioengineering*, *110*(6), 1735–1747. https://doi.org/10.1002/BIT.24826

Duarte, T. M., Carinhas, N., Barreiro, L. C., Carrondo, M. J. T., Alves, P. M., & Teixeira, A. P. (2014). Metabolic responses of CHO cells to limitation of key amino acids. *Biotechnology and Bioengineering*, *111*(10), 2095–2106. https://doi.org/10.1002/BIT.25266

Egea, J. A., Henriques, D., Cokelaer, T., Villaverde, A. F., MacNamara, A., Danciu, D. P., Banga, J. R., & Saez-Rodriguez, J. (2014). MEIGO: An open-source software suite based on metaheuristics for global optimization in systems biology and bioinformatics. *BMC Bioinformatics*, *15*(1), 1–9. https://doi.org/10.1186/1471-2105-15-136/FIGURES/8

Facco, P., Zomer, S., Rowland-Jones, R. C., Marsh, D., Diaz-Fernandez, P., Finka, G., Bezzo, F., & Barolo, M. (2020). Using data analytics to accelerate biopharmaceutical process scale-up. *Biochemical Engineering Journal*, *164*, 107791. https://doi.org/10.1016/J.BEJ.2020.107791

Gaughan, C. L. (2016). The present state of the art in expression, production and characterization of monoclonal antibodies. *Molecular Diversity*, *20*(1), 255–270. https://doi.org/10.1007/S11030-015-9625-Z

Geladi, P., & Kowalski, B. R. (1986). Partial least-squares regression: a tutorial. *Analytica Chimica Acta*, *185*(C), 1–17. https://doi.org/10.1016/0003-2670(86)80028-9

Genzel, Y., Ritter, J. B., König, S., Alt, R., & Reichl, U. (2005). Substitution of glutamine by pyruvate to reduce ammonia formation and growth inhibition of mammalian cells. *Biotechnology Progress*, *21*(1), 58–69. https://doi.org/10.1021/BP049827D

Ghorbaniaghdam, A., Henry, O., & Jolicoeur, M. (2013). A kinetic-metabolic model based on cell energetic state: study of CHO cell behavior under Na-butyrate stimulation. *Bioprocess and Biosystems Engineering*, *36*(4), 469–487. https://doi.org/10.1007/S00449-012-0804-3

Godfrey, K. (1999). Identification of parametric models from experimental data [Book Review]. *IEEE Transactions on Automatic Control*, *44*(12), 2321–2322. https://www.academia.edu/58857827/Identification_of_parametric_models_from_experimental_data_Communications_and_Control_Engineering_Series

González-Leal, I. J., Carrillo-Cocom, L. M., Ramírez-Medrano, A., López-Pacheco, F., Bulnes-Abundis, D., Webb-Vargas, Y., & Alvarez, M. M. (2011). Use of a Plackett-Burman statistical design to determine the effect of selected amino acids on monoclonal antibody production in CHO cells. *Biotechnology Progress*, *27*(6), 1709–1717. https://doi.org/10.1002/BTPR.674

Grindley, J. F., Payton, M. A., Van de Pol, H., & Hardy, K. G. (1988). Conversion of Glucose to 2-Keto-l-Gulonate, an Intermediate in l-Ascorbate Synthesis, by a Recombinant Strain of Erwinia citreus. *Applied and Environmental Microbiology*, *54*(7), 1770–1775. https://doi.org/10.1128/AEM.54.7.1770-1775.1988

Homma, T., & Saltelli, A. (1996). Importance measures in global sensitivity analysis of nonlinear models. *Reliability Engineering & System Safety*, *52*(1), 1–17. https://doi.org/10.1016/0951-8320(96)00002-6

Jang, J. D., & Barford, J. P. (2000). An unstructured kinetic model of macromolecular metabolism in batch and fed-batch cultures of hybridoma cells producing monoclonal antibody. *Biochemical Engineering Journal*, *4*(2), 153–168. https://doi.org/10.1016/S1369-703X(99)00041-8

Jimenez del Val, I., Fan, Y., & Weilguny, D. (2016). Dynamics of immature mAb glycoform secretion during CHO cell culture: An integrated modelling framework. *Biotechnology Journal*, *11*(5), 610–623. https://doi.org/10.1002/BIOT.201400663

Kontoravdi, C. (2006). *An Integrated Modelling/Experimental Framework for Protein-Producing Cell Cultures*.

Kontoravdi, C., Pistikopoulos, E. N., & Mantalaris, A. (2010a). Systematic development of predictive mathematical models for animal cell cultures. *Computers & Chemical Engineering*, *34*(8), 1192–1198. https://doi.org/10.1016/J.COMPCHEMENG.2010.03.012

Kontoravdi, C., Pistikopoulos, E. N., & Mantalaris, A. (2010b). Systematic development of predictive mathematical models for animal cell cultures. *Computers & Chemical Engineering*, *34*(8), 1192–1198. https://doi.org/10.1016/J.COMPCHEMENG.2010.03.012

Krampe, B., & Al-Rubeai, M. (2010). Cell death in mammalian cell culture: molecular mechanisms and cell line engineering strategies. *Cytotechnology*, *62*(3), 175. https://doi.org/10.1007/S10616-010-9274-0

Kyriakopoulos, S., Ang, K. S., Lakshmanan, M., Huang, Z., Yoon, S., Gunawan, R., & Lee, D. Y. (2018). Kinetic Modeling of Mammalian Cell Culture Bioprocessing: The Quest to

Advance Biomanufacturing. *Biotechnology Journal*, *13*(3). https://doi.org/10.1002/BIOT.201700229

Li, F., Vijayasankaran, N., Shen, A., Kiss, R., & Amanullah, A. (2010). Cell culture processes for monoclonal antibody production. *MAbs*, *2*(5), 466. https://doi.org/10.4161/MABS.2.5.12720

Liu, H., Chen, W., & Sudjianto, A. (2006). Relative Entropy Based Method for Probabilistic Sensitivity Analysis in Engineering Design. *Journal of Mechanical Design*, *128*(2), 326–336. https://doi.org/10.1115/1.2159025

Liu, M., Wang, J., Tang, H., Fan, L., Zhao, L., Wang, H. Bin, Zhou, Y., & Tan, W. S. (2018). Cell culture medium supplemented with taurine decreases basic charge variant levels of a monoclonal antibody. *Biotechnology Letters*, *40*(11–12), 1487–1493. https://doi.org/10.1007/S10529-018-2606-4

Lu, R. M., Hwang, Y. C., Liu, I. J., Lee, C. C., Tsai, H. Z., Li, H. J., & Wu, H. C. (2020). Development of therapeutic antibodies for the treatment of diseases. *Journal of Biomedical Science 2020 27:1*, *27*(1), 1–30. https://doi.org/10.1186/S12929-019-0592-Z

McKay, M. D., Beckman, R. J., & Conover, W. J. (1979). Comparison of three methods for selecting values of input variables in the analysis of output from a computer code. *Technometrics*, *21*(2), 239–245. https://doi.org/10.1080/00401706.1979.10489755

Morris, M. D. (1991). *Factorial Sampling Plans for Preliminary Computational Experiments*. *33*(2), 161–174.

Nomikos, P., & MacGregor, J. F. (1994). Monitoring batch processes using multiway principal component analysis. *AIChE Journal*, *40*(8), 1361–1375. https://doi.org/10.1002/AIC.690400809

Nomikos, P., & MacGregor, J. F. (1995a). Multivariate SPC charts for monitoring batch processes. *Technometrics*, *37*(1), 41–59. https://doi.org/10.1080/00401706.1995.10485888

Nomikos, P., & MacGregor, J. F. (1995b). Multi-way partial least squares in monitoring batch processes. *Chemometrics and Intelligent Laboratory Systems*, *30*(1), 97–108. https://doi.org/10.1016/0169-7439(95)00043-7

Phillips, B. J., James, T. E. B., & Gangoli, S. D. (1982). Genotoxicity studies of di(2-ethylhexyl)phthalate and its metabolites in CHO cells. *Mutation Research*, *102*(3), 297–304. https://doi.org/10.1016/0165-1218(82)90139-2

Pianosi, F., Sarrazin, F., & Wagener, T. (2015). A Matlab toolbox for Global Sensitivity Analysis. *Environmental Modelling & Software*, *70*, 80–85. https://doi.org/10.1016/J.ENVSOFT.2015.04.009

Pianosi, F., & Wagener, T. (2015). A simple and efficient method for global sensitivity analysis based on cumulative distribution functions. *Environmental Modelling & Software*, *67*, 1–11. https://doi.org/10.1016/J.ENVSOFT.2015.01.004

Pörtner, R., & Schäfer, T. (1996). Modelling hybridoma cell growth and metabolism--a comparison of selected models and data. *Journal of Biotechnology*, *49*(1–3), 119–135. https://doi.org/10.1016/0168-1656(96)01535-0

Quinteros, D. A., Bermúdez, J. M., Ravetti, S., Cid, A., Allemandi, D. A., & Palma, S. D. (2017). Therapeutic use of monoclonal antibodies: general aspects and challenges for drug delivery. *Nanostructures for Drug Delivery*, 807. https://doi.org/10.1016/B978-0-323-46143-6.00025-7

Rader, R. A. (2008). (Re)defining biopharmaceutical. *Nature Biotechnology*, *26*(7), 743–751. https://doi.org/10.1038/NBT0708-743

Rajalahti, T., Arneberg, R., Berven, F. S., Myhr, K. M., Ulvik, R. J., & Kvalheim, O. M. (2009). Biomarker discovery in mass spectral profiles by means of selectivity ratio plot. *Chemometrics and Intelligent Laboratory Systems*, *95*(1), 35–48. https://doi.org/10.1016/J.CHEMOLAB.2008.08.004

Reel, P. S., Reel, S., Pearson, E., Trucco, E., & Jefferson, E. (2021). Using machine learning approaches for multi-omics data analysis: A review. *Biotechnology Advances*, *49*. https://doi.org/10.1016/J.BIOTECHADV.2021.107739

Saltelli, A., Annoni, P., Azzini, I., Campolongo, F., Ratto, M., & Tarantola, S. (2010). Variance based sensitivity analysis of model output. Design and estimator for the total sensitivity index. *Computer Physics Communications*, *181*, 259–270. https://doi.org/10.1016/j.cpc.2009.09.018

Saltelli, A., Ratto, M., Andres, T., Campolongo, F., Cariboni, J., Gatelli, D., Saisana, M., & Tarantola, S. (2008a). Introduction to Sensitivity Analysis. *Global Sensitivity Analysis. The Primer*, 1–51. https://doi.org/10.1002/9780470725184.CH1

Saltelli, A., Ratto, M., Andres, T., Campolongo, F., Cariboni, J., Gatelli, D., Saisana, M., & Tarantola, S. (2008b). Variance-Based Methods. *Global Sensitivity Analysis. The Primer*, 155–182. https://doi.org/10.1002/9780470725184.CH4

Schneider, M., Marison, I. W., & Von Stockar, U. (1996). The importance of ammonia in mammalian cell culture. *Journal of Biotechnology*, *46*(3), 161–185. https://doi.org/10.1016/0168-1656(95)00196-4

Scott, & al. (2000). *Single amino acid (arginine)deprivation:rapid and selective death of cultured transformed and malignant cells*. https://doi.org/10.1054/bjoc.2000.1353

Takagi, Y., Kikuchi, T., Wada, R., & Omasa, T. (2017). The enhancement of antibody concentration and achievement of high cell density CHO cell cultivation by adding nucleoside. *Cytotechnology*, *69*(3), 511–521. https://doi.org/10.1007/S10616-017-0066-7

Troyanskaya, O., Cantor, M., Sherlock, G., Brown, P., Hastie, T., Tibshirani, R., Botstein, D., & Altman, R. B. (2001). Missing value estimation methods for DNA microarrays. *BIOINFORMATICS*, *17*(6), 520–525. http://smi-web.

Tsao, Y. S., Cardoso, A. G., Condon, R. G. G., Voloch, M., Lio, P., Lagos, J. C., Kearns, B. G., & Liu, Z. (2005). Monitoring Chinese hamster ovary cell culture by the analysis of glucose and lactate metabolism. *Journal of Biotechnology*, *118*(3), 316–327. https://doi.org/10.1016/J.JBIOTEC.2005.05.016

Villaverde, A. F., Barreiro, A., & Papachristodoulou, A. (2016). Structural Identifiability of Dynamic Systems Biology Models. *PLOS Computational Biology*, *12*(10), e1005153. https://doi.org/10.1371/JOURNAL.PCBI.1005153

Walsh, G., & Walsh, E. (2022). Biopharmaceutical benchmarks 2022. *Nature Biotechnology 2022 40:12*, *40*(12), 1722–1760. https://doi.org/10.1038/s41587-022-01582-x

Wang, T., Gnanaprakasam, J. N. R., Chen, X., Kang, S., Xu, X., Sun, H., Liu, L., Rodgers, H., Miller, E., Cassel, T. A., Sun, Q., Vicente-Muñoz, S., Warmoes, M. O., Lin, P., Piedra-Quintero, Z. L., Guerau-de-Arellano, M., Cassady, K. A., Zheng, S. G., Yang, J., … Wang, R. (2020). Inosine is an alternative carbon source for CD8+-T-cell function under glucose restriction. *Nature Metabolism*, *2*(7), 635–647. https://doi.org/10.1038/S42255-020-0219-4

Wise, B. M., & Gallagher, N. B. (1996). The process chemometrics approach to process monitoring and fault detection. *Journal of Process Control*, *6*(6), 329–348. https://doi.org/10.1016/0959-1524(96)00009-1

Wold, S. (1978). Cross-Validatory Estimation of the Number of Components in Factor and Principal Components Models. *Technometrics*, *20*(4), 397. https://doi.org/10.2307/1267639

Wold, S., Sjöström, M., & Eriksson, L. (2001). PLS-regression: a basic tool of chemometrics. *Chemometrics and Intelligent Laboratory Systems*, *58*(2), 109–130. https://doi.org/10.1016/S0169-7439(01)00155-1

Xing, Z., Bishop, N., Leister, K., & Li, Z. J. (2010). Modeling kinetics of a large-scale fed-batch CHO cell culture by Markov chain Monte Carlo method. *Biotechnology Progress*, *26*(1), 208–219. https://doi.org/10.1002/BTPR.284

Xu, W. J., Lin, Y., Mi, C. L., Pang, J. Y., & Wang, T. Y. (2023). Progress in fed-batch culture for recombinant protein production in CHO cells. *Applied Microbiology and Biotechnology 2023 107:4*, *107*(4), 1063–1075. https://doi.org/10.1007/S00253-022-12342-X

Yadav, P., Chauhan, A. K., Singh, R. B., Khan, S., & Halabi, G. (2022). Organic acids: microbial sources, production, and applications. *Functional Foods and Nutraceuticals in Metabolic and Non-Communicable Diseases*, 325–337. https://doi.org/10.1016/B978-0-12-819815-5.00053-7

Young, J. D. (2013). Metabolic flux rewiring in mammalian cell cultures. *Current Opinion in Biotechnology*, *24*(6), 1108–1115. https://doi.org/10.1016/J.COPBIO.2013.04.016

Acknowledgments

At the end of my student career, I think it is right to thank everyone who helped me during this period of my personal life.
First of all, I would like to thank my Thesis advisor (and hopefully my future supervisor), Prof. Pierantonio Facco. Thank you for all the support since the first moment we met and for being always present and helpful. I would also like to thank him for helping me to take an important step forward in my professional career.
A very big thank you goes to Gianmarco Barberi, without his presence everything would have been much more difficult. Your support during this period has been very important, I appreciate the availability that has never been negated to me, and all your valid suggestions. I would like to thank you very much for having been a point of reference.
Thanks to the industrial collaborators at GSK, especially to Mrs. Paloma Diaz Fernandez, for the opportunity to work on stimulating, challenging, and impactful problems.
Un enorme ringraziamento va alla mia Famiglia, Lisa, Emma e Andrea. Grazie per avermi aiutato nei momenti più difficili e per essermi stati sempre vicini. Senza il vostro aiuto e il vostro supporto non sarei mai riuscito a diventare la persona che sono, grazie per aver sempre creduto in me.
Un ringraziamento va a tutti gli amici di questi anni che hanno completato le mie giornate e con risate e momenti di grande amicizia. Un ringraziamento in particolare va a Mattia. Ti ringrazio per essere un amico fidato, unico nel suo genere. Non ti ringrazierò mai abbastanza per avermi dato la possibilità di coltivare questa amicizia sin dalla nostra infanzia.
Ultima, non per importanza, un ringraziamento speciale va a Greta. In questa esperienza sei stata una compagna fantastica, con la quale ho avuto il piacere di condividere gioie e dolori. Grazie per essere sempre riuscita ad essere un faro nei momenti più bui, la tua dote nel farmi tornare s il buonumore è stata (e sarà) sempre, il motivo per cui le emozioni di quel giorno si ripresentano ancora, anche dopo quattro anni. Spero che questi anni assieme possano essere l'inizio di una avventura da scoprire insieme. Tu *sei* per me, il pezzo del Tetris longilineo.